T·L·G
Texts in Logic and Games
Volume 3

Logic and the Foundations of Game and Decision Theory (LOFT 7)

EDITED BY
GIACOMO BONANNO, WIEBE VAN DER HOEK AND
MICHAEL WOOLDRIDGE

AMSTERDAM UNIVERSITY PRESS

LOGIC AND THE FOUNDATIONS OF
GAME AND DECISION THEORY (LOFT 7)

T·L·G
Texts in Logic and Games
Volume 3

Logic and the Foundations of Game and Decision Theory (LOFT 7)

EDITED BY
GIACOMO BONANNO
WIEBE VAN DER HOEK
MICHAEL WOOLDRIDGE

Texts in Logic and Games
Volume 3

AMSTERDAM UNIVERSITY PRESS

Cover design: Maedium, Utrecht

ISBN 978 90 8964 026 0
NUR 918

Table of Contents

Preface

This volume in the *Texts in Logic and Games* series was conceived as a ramification of the seventh conference on *Logic and the Foundations of the Theory of Games and Decisions* (LOFT7), which took place in Liverpool, in July 2006.[1]

The LOFT conferences have been a regular biannual event since 1994. The first conference was hosted by the Centre International de Recherches Mathematiques in Marseille (France), the next four took place at the International Centre for Economic Research in Torino (Italy), the sixth conference was hosted by the Graduate School of Management in Leipzig (Germany) and the most recent one took place at the University of Liverpool (United Kingdom).[2]

The LOFT conferences are interdisciplinary events that bring together researchers from a variety of fields: computer science, economics, game theory, linguistics, logic, multi-agent systems, psychology, philosophy, social choice and statistics. In its original conception, LOFT had as its central theme the application of logic, in particular modal epistemic logic, to foundational issues in the theory of games and individual decision-making. Epistemic considerations have been central to game theory for a long time. The

[1] The conference was organized by the editors of this volume with the assistance of a program committee consisting of Thomas Ågotnes, Johan van Benthem, Adam Brandenburger, Hans van Ditmarsch, Jelle Gerbrandy, Wojtek Jamroga, Hannes Leitgeb, Benedikt Löwe, Marc Pauly, Andrés Perea, Gabriella Pigozzi, Wlodek Rabinowicz, Hans Rott, and Krister Segerberg.

[2] Collections of papers from previous LOFT conferences can be found in a special issue of *Theory and Decision* (Vol. 37, 1994, edited by M. Bacharach and P. Mongin), the volume *Epistemic logic and the theory of games and decisions* (edited by M. Bacharach, L.-A. Gérard-Varet, P. Mongin and H. Shin and published by Kluwer Academic, 1997), two special issues of *Mathematical Social Sciences* (Vols. 36 and 38, 1998, edited by G. Bonanno, M. Kaneko and P. Mongin), two special issues of *Bulletin of Economic Research* (Vol. 53, 2001 and Vol. 54, 2002, edited by G. Bonanno and W. van der Hoek), a special issue of *Research in Economics,* (Vol. 57, 2003, edited by G. Bonanno and W. van der Hoek), a special issue of *Knowledge, Rationality and Action* (part of *Synthese*, Vol. 147, 2005, edited by G. Bonanno) and the volume *Proceedings of the 7th Conference on Logic and the Foundations of Game and Decision Theory* (edited by G. Bonanno, W. van der Hoek and M. Wooldridge, University of Liverpool, 2006).

expression "interactive epistemology" has been used in the game-theory literature to refer to the analysis of decision making by agents involved in a strategic interaction, when these agents recognize each other's intelligence and rationality. What is relatively new is the realization that the tools and methodologies that were used in game theory are closely related to those already used in other fields, notably computer science and philosophy. Modal logic turned out to be the common language that made it possible to bring together different professional communities. It became apparent that the insights gained and the methodologies employed in one field could benefit researchers in other fields. Indeed, new and active areas of research have sprung from the interdisciplinary exposure provided by the LOFT conferences.[3]

Over time the scope of the LOFT conference has broadened to encompass a wider range of topics, while maintaining its focus on the general issue of rationality and agency. Topics that have fallen within the LOFT umbrella include epistemic and temporal logic, theories of information processing and belief revision, models of bounded rationality, non-monotonic reasoning, theories of learning and evolution, mental models, etc.

The papers collected in this volume reflect the general interests and interdisciplinary scope of the LOFT conferences.

The paper by Alexandru Baltag and Sonja Smets falls within the recent literature that deals with belief revision and update within the Dynamic Epistemic Logic paradigm. The authors develop a notion of doxastic action general enough to cover many examples of multi-agent communication actions encountered in the literature, but also flexible enough to deal with both static and dynamic belief revision. They discuss several epistemic notions: knowledge, belief and conditional belief. For the latter they distinguish between the statement 'if informed that P, the agent would believe that Q *was* the case (before the learning)' and the statement 'if informed that P, the agent would come to believe that Q *is* the case (in the world after the learning)'. They also study a "safe belief" operator meant to express a weak notion of "defeasible knowledge": it is belief that is persistent under revision with any true information. Baltag and Smets provide a complete axiomatization of the logic of conditional belief, knowledge and safe belief. In the second part of the paper the authors discuss dynamic belief revision in the context of action models.

The paper by Giacomo Bonanno deals with the question of what choices are compatible with rationality of the players and common belief of rationality. He takes a syntactic approach and defines rationality axiomatically.

[3] There is substantial overlap between the LOFT community and the community of researchers who are active in another regular, biannual event, namely the conferences on Theoretical Aspects of Rationality and Knowledge (TARK).

Furthermore, he does not assume von Neumann-Morgenstern payoffs but merely ordinal payoffs, thus aiming for a more general theory of rationality in games. The author considers two axioms. The first says that a player is irrational if she chooses a particular strategy while believing that another strategy of hers is better. He shows that common *belief* of this weak notion of rationality characterizes the iterated deletion of pure strategies that are strictly dominated by another pure strategy. The second axiom says that a player is irrational if she chooses a particular strategy while believing that a different strategy is at least as good and she considers it possible that this alternative strategy is actually better than the chosen one. The author shows that common *knowledge* of this stronger notion of rationality characterizes the iterated deletion procedure introduced by Stalnaker (1994), restricted—once again—to pure strategies.

The paper by Hans van Ditmarsch and Barteld Kooi investigates a dynamic logic describing "epistemic events" that may change both the agents' information (or beliefs) and what the authors call "the ontic facts" of the world (that is, objective, non-epistemic statements about the world). A sound and complete axiomatization is provided. Some original and interesting semantic results are also proved, in particular the fact that any model change can be simulated by "epistemic events", and thus any consistent goal can be achieved by performing some such event. The authors illustrate their results in several examples, including card games and logical puzzles.

The paper by Wiebe van der Hoek, Mark Roberts and Michael Wooldridge extends the authors' previous work on Alternating-time Temporal Logic and its ramifications. They extend it by introducing the notion of a legally possible strategy, that they oppose to a physically possible strategy, and define social belief as truth in all states that are (1) possible for the agent, and (2) are obtained from the initial state by a legally possible strategy. They use this framework to reason about social laws. In a system with social laws, every agent is supposed to refrain from performing certain forbidden actions. Rather than assuming that all agents abide by the law, the authors consider what happens if certain agents act socially, while others do not. In particular, they focus on the agents' strategic abilities under such mixed conditions.

The paper by Alexander Nittka and Richard Booth deals with the traditional "static" belief revision setting, but with a different twist: rather than answering the question of how an agent should rationally change his beliefs in the light of new information, they address the question of what one can say about an agent who is observed in a belief change process. That is, the authors study the problem of how to make inferences about an agent's beliefs based on observation of how that agent responded to a

sequence of revision inputs over time. They start by reviewing some earlier results for the case where the observation is complete in the sense that (1) the logical content of all formulas appearing in the observation is known, and (2) all revision inputs received by the agent during the observed period are recorded in the observation. They then provide new results for the more general case where information in the observation might be distorted due to noise or because some revision inputs are missing altogether. Their analysis is based on the assumption that the agent employs a specific, but plausible, belief revision framework when incorporating new information.

The paper by R. Ramanujam and Sunil Simon deals with the most important notion of non-cooperative game, namely extensive game. Extensive games provide a richer description of interactive situations than strategic-form games in that they make the order of moves and the information available to a player when it is his turn to move explicit. A strategy for a player in an extensive game associates with every information set of that player a choice at that information set. The authors observe that the game position (or information set) may be only partially known, in terms of properties that the player can test for. Thus—they argue—strategies can be thought of as programs, built up systematically from atomic decisions like *if* b *then* a where b is a condition checked by the player to hold (at some game position) and a is a move available to the player at that position. This leads them to propose a logical structure for strategies, where one can reason with assertions of the form "(partial) strategy σ ensures the (intermediate) condition α". They present an axiomatization for the logic and prove its completeness.

The paper by Giacomo Sillari contributes to the very recent and fast growing literature on the notion of (un)awareness. An open problem in this literature has been how to model the state of mind of an individual who realizes that he may be unaware of something, that is, the problem of formalizing the notion of "awareness of unawareness". Sillari offers a solution to this problem using a new system of first-order epistemic logic with awareness. He also offers a philosophical analysis of awareness structures and proves that a certain fragment of the first-order epistemic language with awareness operators is decidable.

The papers went through a thorough refereeing and editorial process. The editors would like to thank the many referees who provided invaluable help and the authors for their cooperation during the revision stage.

Davis, CA & Liverpool G.B. W.v.d.H. M.W.

A Qualitative Theory of Dynamic Interactive Belief Revision

Alexandru Baltag[1]
Sonja Smets[2,3]

[1] Computing Laboratory
Oxford University
Oxford OX1 3QD, United Kingdom

[2] Center for Logic and Philosophy of Science
Vrije Universiteit Brussel
Brussels, B1050, Belgium

[3] IEG Research Group on the Philosophy of Information
Oxford University
Oxford OX1 3QD, United Kingdom

baltag@comlab.ox.ac.uk, sonsmets@vub.ac.be

Abstract

We present a logical setting that incorporates a belief-revision mechanism within Dynamic-Epistemic logic. As the "static" basis for belief revision, we use *epistemic plausibility models*, together with a modal language based on *two epistemic operators*: a *"knowledge" modality* K (the standard S5, fully introspective, notion), and a *"safe belief" modality* $\Box$ ("weak", non-negatively-introspective, notion, capturing a version of Lehrer's "indefeasible knowledge"). To deal with "dynamic" belief revision, we introduce *action plausibility models*, representing various types of "doxastic events". Action models "act" on state models via a modified update product operation: the *"Action-Priority" Update*. This is the natural dynamic generalization of AGM revision, giving priority to the incoming information (i.e., to "actions") over prior beliefs. We completely axiomatize this logic, and show how our update mechanism can "simulate", in a uniform manner, *many different belief-revision policies*.

1 Introduction

This paper contributes to the recent and on-going work in the logical community [2, 14, 24, 8, 10, 9, 7] on dealing with mechanisms for belief revision and update within the Dynamic-Epistemic Logic (DEL) paradigm. DEL originates in the work of Gerbrandy and Groeneveld [30, 29], anticipated by Plaza in [44], and further developed by numerous authors

Giacomo Bonanno, Wiebe van der Hoek, Michael Wooldridge (eds.). *Logic and the Foundations of Game and Decision Theory (LOFT 7)*. Texts in Logic and Games 3, Amsterdam

[6, 31, 22, 4, 23, 39, 5, 15, 16] etc. In its standard incarnation, as presented e.g., in the recent textbook [25], the DEL approach is particularly well fit to deal with *complex multi-agent learning actions* by which groups of interactive agents update their beliefs (including *higher-level beliefs* about the others' beliefs), *as long as the newly received information is consistent with the agents' prior beliefs*. On the other hand, the classical AGM theory and its more recent extensions have been very successful in dealing with the problem of *revising one-agent, first-level (factual) beliefs when they are contradicted by new information*. So it is natural to look for a way to combine these approaches.

We develop here a notion of *doxastic actions*[1], general enough to cover most examples of multi-agent communication actions encountered in the literature, but also flexible enough to deal with *(both static and dynamic) belief revision*, and in particular to *implement various "belief-revision policies" in a unified setting*. Our approach can be seen as a natural extension of the work in [5, 6] on "epistemic actions", incorporating ideas from the AGM theory along the lines pioneered in [2] and [24], but using a *qualitative* approach based on *conditional beliefs*, in the line of [50, 20, 19, 14].

Our paper assumes the general distinction, made in [24, 8, 14], between *"dynamic" and "static" belief revision*. It is usually acknowledged that the classical AGM theory in [1, 28] (and embodied in our setting by the *conditional belief* operators $B_a^P Q$) is indeed "static", in the sense that it captures *the agent's changing beliefs about an unchanging world*. But in fact, when we take into account all the higher-level beliefs, the "world" (that these higher-level beliefs are about) includes all agent's (real) beliefs.[2] Thus, such a world is *always changed by our changes of beliefs*! So we can better understand a belief conditional on P as capturing the agent's beliefs *after revising with* P about the state of the world *before the revision*: the statement $B_a^P Q$ says that, *if agent a would learn P, then she would come to believe that Q was the case (before the learning)*. In contrast, "dynamic" belief revision uses dynamic modalities to capture the agent's revised beliefs about the world *as it is after revision*: $[!\,P]B_a Q$ says that *after learning P, agent a would come to believe that Q is the case (in the world after the learning)*. The standard alternative [37] to the AGM theory calls this *belief update*, but like the AGM approach, it only deals with "first-level" (factual) beliefs from a non-modal perspective, neglecting any higher-order "beliefs about beliefs". As a result, *it completely misses the changes induced* (in our own or the other agents' epistemic-doxastic states) *by the learning actions themselves* (e.g., the learning of a Moore sentence, see Section 3). This

[1] Or "doxastic events", in the terminology of [14].

[2] To verify that a higher-level belief about another belief is "true" we need to check the content of that higher-level belief (i.e., the existence of the second, lower-level belief) against the "real world". So the real world has to include the agent's beliefs.

is reflected in the acceptance in [37] of the AGM "Success Axiom": in dynamic notation, this is the axiom $[!P]B_aP$ (which cannot accommodate Moore sentences). Instead, the authors of [37] exclusively concentrate on the possible changes of (ontic) facts that may have occurred during our learning (but *not due to our learning*). In contrast, our approach to belief update (following the DEL tradition) may be thought of as "dual" to the one in [37]: we completely neglect here the ontic changes[3], considering only the changes induced by *"purely doxastic" actions* (learning by observation, communication, etc.).

Our formalism for "static" revision can best be understood as a modal-logic implementation of the well-known view of belief revision in terms of *conditional reasoning* [50, 52]. In [8] and [10], we introduced two equivalent semantic settings for conditional beliefs in a multi-agent epistemic context (*conditional doxastic models* and *epistemic plausibility models*), taking the first setting as the basic one. Here, we adopt the second setting, which is closer to the standard semantic structures used in the literature on modeling belief revision [34, 49, 52, 27, 19, 14, 17]. We use this setting to define notions of *knowledge* K_aP, *belief* B_aP and *conditional belief* B_a^QP. Our concept of "knowledge" is the standard S5-notion, partition-based and fully introspective, that is commonly used in Computer Science and Economics, and is sometimes known as "Aumann knowledge", as a reference to [3]. The conditional belief operator is a way to "internalize", in a sense, the "static" (AGM) belief revision within a modal framework: saying that, at state s, agent a believes P conditional on Q is a way of saying that Q belongs to a's revised "theory" (capturing her revised beliefs) after revision with P (of a's current theory/beliefs) at state s. Our conditional formulation of "static" belief revision is close to the one in [50, 47, 19, 20, 45]. As in [19], the preference relation is assumed to be well-preordered; as a result, the logic CDL of conditional beliefs is equivalent to the strongest system in [19].

We also consider other modalities, capturing other "doxastic attitudes" than just knowledge and conditional belief. The most important such notion expresses a form of "weak (non-introspective) knowledge" $\Box_aP$, first introduced by Stalnaker in his modal formalization [50, 52] of Lehrer's *defeasibility analysis of knowledge* [40, 41]. We call this notion *safe belief*, to distinguish it from our (Aumann-type) concept of knowledge. Safe belief can be understood as belief that is *persistent under revision with any true information*. We use this notion to give a new solution to the so-called "Paradox of the Perfect Believer". We also solve the open problem posed in [19], by providing a *complete axiomatization of the "static" logic* $K\Box$ *of conditional belief, knowledge and safe belief*. In a forthcoming paper, we

[3] But our approach can be easily modified to incorporate ontic changes, along the lines of [15].

apply the concept of safe belief to Game Theory, improving on Aumann's epistemic analysis of backwards induction in games of perfect information.

Moving thus on to *dynamic belief revision*, the first thing to note is that (unlike the case of "static" revision), *the doxastic features of the actual "triggering event"* that induced the belief change *are essential* for understanding this change (as a "dynamic revision", i.e., in terms of the revised beliefs about the state of the world after revision). For instance, our beliefs about *the current situation after* hearing a *public* announcement (say, of some *factual* information, denoted by an atomic sentence p) are different from our beliefs after receiving a *fully private* announcement with the same content p. Indeed, in the public case, we come to believe that p is now *common knowledge* (or at least *common belief*). While, in the private case, we come to believe that the content of the announcement forms now our *secret knowledge*. So the agent's *beliefs about the learning actions* in which she is currently engaged affect the way she updates her previous beliefs.

This distinction is irrelevant for "static" revision, since e.g., in both cases above (public as well as private announcement) we learn the same thing about the situation that existed *before the learning*: our beliefs about that past situation will change in the same way in both cases. More generally, our beliefs about the "triggering action" are irrelevant, as far as our "static" revision is concerned. This explains a fact observed in [14], namely that by and large, the standard literature on belief revision (or belief update) *does not usually make explicit the doxastic events that "trigger" the belief change* (dealing instead only with types of abstract operations on beliefs, such as update, revision and contraction etc). The reason for this lies in the "static" character of AGM revision, as well as its restriction (shared with the "updates" of [37]) to one-agent, first-level, factual beliefs.

A "truly dynamic" logic of belief revision has to be able to capture the *doxastic-epistemic features (e.g., publicity, complete privacy etc.) of specific "learning events"*. We need to be able to model the agents' "dynamic beliefs", i.e., their *beliefs about the learning action itself*: the *appearance* of this action (while it is happening) to each of the agents. In [5], it was argued that a natural way to do this is to use *the same type of formalism that was used to model "static" beliefs*: *epistemic actions should be modeled in essentially the same way as epistemic states*; and this common setting was taken there to be given by *epistemic Kripke models*.

A similar move is made here in the context of our richer doxastic-plausibility structures, by introducing *plausibility pre-orders on actions* and developing a notion of "action plausibility models", that extends the "epistemic action models" from [5], along similar lines to (but without the quantitative features of) the work in [2, 24].

Extending to (pre)ordered models the corresponding notion from [5], we

introduce an operation of *product update* of such models, based on the *anti-lexicographic order* on the product of the state model with the action model. The simplest and most natural way to define a connected pre-order on a Cartesian product from connected pre-orders on each of the components is to use either the *lexicographic* or the *anti-lexicographic* order. Our choice is the second, which we regard as the *natural generalization of the AGM theory*, giving *priority to incoming information* (i.e., to "actions" in our sense). This can also be thought of as a generalization of the so-called *"maximal-Spohn" revision*. We call this type of update rule the *"Action-Priority" Update*. The intuition is that the beliefs encoded in the action model express *the "incoming" changes of belief*, while the state model only captures that *past beliefs*. One could say that the new "beliefs about actions" are *acting* on the prior "beliefs about states", producing the updated (posterior) beliefs. This is embedded in the Motto of Section 3.1: "*beliefs about changes encode (and induce) changes of beliefs*".

By abstracting away from the quantitative details of the plausibility maps when considering the associated *dynamic logic*, our approach to dynamic belief revision is in the spirit of the one in [14]: instead of using "graded belief" operators as in e.g., [2, 24], or probabilistic modal logic as in [39], both our account and the one in [14] concentrate on the simple, qualitative language of *conditional beliefs, knowledge and action modalities* (to which we add here the *safe belief* operator). As a consequence, we obtain *simple, elegant, general logical laws of dynamic belief revision*, as natural generalizations of the ones in [14]. These "reduction laws" give a *complete axiomatization of the logic of doxastic actions*, "reducing" it to the "static" logic $K\Box$. Compared both to our older axiomatization in [10] and to the system in [2], one can easily see that the introduction of the safe belief operator leads to a major simplification of the reduction laws.

Our qualitative logical setting (in this paper and in [8, 10, 9]), as well as van Benthem's closely related setting in [14], are conceptually very different from the more "quantitative" approaches to dynamic belief revision taken by Aucher, van Ditmarsch and Labuschagne [2, 24, 26], approaches based on "degrees of belief" given by ordinal plausibility functions. This is not just a matter of interpretation, but it makes a difference for the choice of dynamic revision operators. Indeed, the update mechanisms proposed in [49, 2, 24] are essentially quantitative, using various binary functions in transfinite ordinal arithmetic, in order to compute the degree of belief of the output-states in terms of the degrees of the input-states and the degrees of the actions. This leads to an increase in complexity, both in the computation of updates and in the corresponding logical systems. Moreover, there seems to be no canonical choice for the arithmetical formula for updates, various authors proposing various formulas. No clear intuitive justification

is provided to any of these formulas, and we see no transparent reason to prefer one to the others. In contrast, classical (AGM) belief revision theory is a qualitative theory, based on natural, intuitive postulates, of great generality and simplicity.

Our approach retains this qualitative flavor of the AGM theory, and aims to build a theory of "dynamic" belief revision of equal simplicity and naturality as the classical "static" account. Moreover (unlike the AGM theory), it aims to provide a *"canonical"* choice for a dynamic revision operator, given by our "Action Priority" update. This notion is a *purely qualitative one*[4], based on a *simple, natural relational definition*. From a *formal point of view*, one might see our choice of the anti-lexicographic order as *just one of the many possible options* for developing a belief-revision-friendly notion of update. As already mentioned, it is a generalization of the "maximal-Spohn" revision, already explored in [24] and [2], among many other possible formulas for combining the "degrees of belief" of actions and states. But here we justify our option, arguing that our *qualitative interpretation of the plausibility order makes this the only reasonable choice*.

It may seem that by making this choice, we have confined ourselves to *only one of the bewildering multitude of "belief revision policies"* proposed in the literature [49, 45, 48, 2, 24, 17, 14]. But, as argued below, *this apparent limitation is not so limiting after all*, but can instead be regarded as an *advantage*: the power of the "action model" approach is reflected in the fact that *many different belief revision policies* can be recovered as *instances of the same type of update operation*. In this sense, our approach can be seen as a *change of perspective*: the diversity of possible revision policies is replaced by the diversity of possible action models; the differences are now viewed as *differences in input, rather than having different "programs"*. For a computer scientist, this resembles "Currying" in lambda-calculus: if every "operation" is encoded as an input-term, then *one operation* (functional application) *can simulate all operations*.[5] In a sense, this is nothing but the idea of Turing's universal machine, which underlies universal computation.

The title of our paper is a paraphrase of Oliver Board's "Dynamic Interactive Epistemology" [19], itself a paraphrase of the title ("Interactive Epistemology") of a famous paper by Aumann [3]. We interpret the word "interactive" as referring to the *multiplicity of agents* and the *possibility*

[4] One could argue that our plausibility pre-order relation is equivalent to a quantitative notion (of ordinal degrees of plausibility, such as [49]), but unlike in [2, 24] the way belief update is defined in our account does not make any use of the ordinal "arithmetic" of these degrees.

[5] Note that, as in untyped lambda-calculus, the input-term encoding the operation (i.e., our "action model") and the "static" input-term to be operated upon (i.e., the "state model") are essentially *of the same type*: epistemic plausibility models for the same language (and for the same set of agents).

of communication. Observe that "interactive" does not necessarily imply "dynamic": indeed, Board and Stalnaker consider Aumann's notion to be "static" (since it doesn't accommodate any non-trivial belief revision). But even Board's logic, as well as Stalnaker's [52], are "static" in our sense: they cannot directly capture the effect of learning *actions* (but can only express "static" conditional beliefs). In contrast, our DEL-based approach has all the "dynamic" features and advantages of DEL: in addition to "simulating" a range of individual belief-revision policies, it can deal with an even wider range of *complex types of multi-agent learning and communication actions*. We thus think it is realistic to expect that, *within its own natural limits*[6], our Action-Priority Update Rule could play the role of a *"universal machine" for qualitative dynamic interactive belief-revision*.

2 "Static" Belief Revision

Using the terminology in [14, 8, 10, 9, 11], "static" belief revision is about *pre-encoding potential belief revisions as conditional beliefs*. A conditional belief statement $B_a^P Q$ can be thought of as expressing a "doxastic predisposition" or a "plan of doxastic action": the agent is determined to believe that Q was the case, if he learnt that P was the case. The semantics for conditional beliefs is usually given in terms of plausibility models (or equivalent notions, e.g., "spheres", "onions", ordinal functions etc.) As we shall see, both *(Aumann, S5-like) knowledge* and *simple (unconditional) belief* can be defined in terms of conditional belief, which itself could be defined in terms of a *unary belief-revision operator*: $*_a P$ captures *all the revised beliefs* of agent a after revising (her current beliefs) with P.

In addition, we introduce a *safe belief* operator $\Box_a P$, meant to express a weak notion of "defeasible knowledge" (obeying the laws of the modal logic $S4.3$). This concept was defined in [52, 19] using a higher-order semantics (quantifying over conditional beliefs). But this is in fact equivalent to a first-order definition, as the Kripke modality for the (converse) plausibility relation. This observation greatly simplifies the task of completely axiomatizing the logic of safe belief and conditional beliefs: indeed, our proof system $K\Box$ below is a solution to the open problem posed in [19].

2.1 Plausibility models: The single agent case

To warm up, we consider first the case of only *one agent*, a case which fits well with the standard models for belief revision.

A *single-agent plausibility frame* is a structure $(S, \leq)$, consisting of a set S of "states" and a "well-preorder" $\leq$, i.e., a reflexive, transitive binary

[6] E.g., our update cannot deal with "forgetful" agents, since "perfect recall" is in-built. But finding out what exactly are the "natural limits" of our approach is for now an open problem.

relation on S such that *every non-empty subset has minimal elements*. Using the notation

$$\mathrm{Min}_{\leq} P := \{s \in P : s \leq s' \text{ for all } s' \in P\}$$

for the set of $\leq$-minimal elements of P, the last condition says that: For every set $P \subseteq S$, if $P \neq \varnothing$ then $\mathrm{Min}_{\leq} P \neq \varnothing$.

The usual reading of $s \leq t$ is that "state s is *at least as plausible* as state t". We keep this reading for now, though we will later get back to it and clarify its meaning. The "minimal states" in $\mathrm{Min}_{\leq} P$ are thus the "most plausible states" satisfying proposition P. As usual, we write $s < t$ iff $s \leq t$ but $t \not\leq s$, for the *"strict" plausibility relation* (s is *more plausible* than t). Similarly, we write $s \cong t$ iff both $s \leq t$ and $t \leq s$, for the *"equi-plausibility"* (or *indifference*) relation (s and t are *equally plausible*).

S-propositions and models. Given an epistemic plausibility frame S, an *S-proposition* is any subset $P \subseteq S$. Intuitively, we say that *a state s satisfies the proposition P* if $s \in P$. Observe that a plausibility frame is just a special case of a *relational frame* (or *Kripke frame*). So, as it is standard for Kripke frames in general, we can define a *plausibility model* to be a structure $\mathbf{S} = (S, \leq, \|\cdot\|)$, consisting of a plausibility frame $(S, \leq)$ together with a valuation map $\|\cdot\| : \Phi \to \mathcal{P}(S)$, mapping every element of a given set Φ of "atomic sentences" into S-propositions.

Interpretation. The elements of S will represent the *possible states* (or "possible worlds") of a system. The atomic sentences $p \in \Phi$ represent *"ontic" (non-doxastic) facts*, that might hold or not in a given state. The valuation tells us which facts hold at which worlds. Finally, the plausibility relations $\leq$ capture the agent's *(conditional) beliefs about the state* of the system; if e.g., the agent was given the information that the state of the system is either s or t, she would believe that the system was in the *most plausible* of the two. So, if $s < t$, the agent would believe the real state was s; if $t < s$, she would believe it was t; otherwise (if $s \cong t$), the agent would be indifferent between the two alternatives: she will not be able to decide to believe any one alternative rather than the other.

Propositional operators, Kripke modalities. For every model $\mathbf{S}$, we have the usual Boolean operations with S-propositions

$$P \wedge Q := P \cap Q, \quad P \vee Q := P \cup Q,$$

$$\neg P := S \setminus P, \quad P \to Q := \neg P \vee Q,$$

as well as Boolean constants $\top_S := S$ and $\bot_S := \varnothing$. Obviously, one also introduces *infinitary* conjunctions and disjunctions. In addition, any binary

relation $R \subseteq S \times S$ on S gives rise to a *Kripke modality* $[R] : \mathcal{P}(S) \to \mathcal{P}(S)$, defined by

$$[R]Q := \{s \in S : \forall t\,(sRt \Rightarrow t \in Q)\}.$$

Accessibility relations for belief, conditional belief and knowledge. To talk about beliefs, we introduce a *doxastic accessibility relation* $\to$, given by

$$s \to t \text{ iff } t \in \mathrm{Min}_{\leq} S\,.$$

We read this as saying that: when the actual state is s, the agent believes that *any* of the states t with $s \to t$ *may be* the actual state. This matches the above interpretation of the preorder: the states believed to be possible are the minimal (i.e., "most plausible") ones.

In order to talk about *conditional beliefs*, we can similarly define a *conditional doxastic accessibility relation* for each S-proposition $P \subseteq S$:

$$s \to^P t \text{ iff } t \in \mathrm{Min}_{\leq} P\,.$$

We read this as saying that: when the actual state is s, if the agent is given the information (that) P (is true at the actual state), then she believes that *any* of the states t with $s \to t$ *may be* the actual state.

Finally, to talk about knowledge, we introduce a relation of *epistemic possibility* (or "indistinguishability") $\sim$. Essentially, this is just the universal relation:

$$s \sim t \text{ iff } s, t \in S\,.$$

So, in single-agent models, *all* the states in S are assumed to be "epistemically possible": the only thing *known* with absolute certainty about the current state is that it belongs to S. This is natural, in the context of a single agent: the states known to be impossible are *irrelevant* from the point of doxastic-epistemic logic, so they can simply be excluded from our model S. (As seen below, this cannot be done in the case of multiple agents!)

Knowledge and (conditional) belief. We define *knowledge* and *(conditional) belief* as the Kripke modalities for the epistemic and (conditional) doxastic accessibility relations:

$$KP := [\sim]P\,,$$

$$BP := [\to]P\,,$$

$$B^Q P := [\to^Q]P\,.$$

We read KP as saying that the (implicit) agent *knows* P. This is "knowledge" in the strong Leibnizian sense of "truth in all possible worlds". We similarly read BP as "P is believed" and $B^Q P$ as "P is believed given (or

conditional on) Q". As for *conditional belief* statements $s \in B^Q P$, we interpret them in the following way: if the actual state is s, then after coming to believe that Q is the case (at this actual state), the agent will believe that P *was* the case (at the same actual state, *before* his change of belief). In other words, conditional beliefs B^Q give descriptions of the agent's *plan* (or *commitments*) about what he will believe about the current state after receiving new (believable) information. To quote Johan van Benthem in [14], conditional beliefs are *"static pre-encodings" of the agent's potential belief changes* in the face of new information.

Discussion on interpretation. Observe that our interpretation of the plausibility relations is *qualitative*, in terms of *conditional beliefs* rather than "degrees of belief": there is no scale of beliefs here, allowing for "intermediary" stages between believing and not believing. Instead, all these beliefs are equally "firm" (*though conditional*): given the condition, something is either believed or not. To repeat, writing $s < t$ is for us just a way to say that: if *given* the information that the state of the system is either s or t, the agent would believe it to be s. So plausibility relations are special cases of conditional belief. This interpretation is based on the following (easily verifiable) equivalence:

$$s < t \text{ iff } s \in B^{\{s,t\}}\{s\} \text{ iff } t \in B^{\{s,t\}}\{s\}.$$

There is nothing quantitative here, no need for us to refer in any way to the "strength" of this agent's belief: though she might have beliefs of unequal strengths, we are not interested here in modeling this quantitative aspect. Instead, we give the agent some information about a state of a virtual system (that it is either s or t) and we ask her a *yes-or-no question* ("Do you believe that virtual state to be s ?"); we write $s < t$ iff the agent's answer is "yes". This is a firm answer, so it expresses a firm belief. "Firm" does not imply "un-revisable" though: if later we reveal to the agent that the state in question was in fact t, she should be able to accept this new information; after all, the agent should be introspective enough to realize that her belief, however firm, was just a belief.

One possible objection against this qualitative interpretation is that our postulate that $\leq$ is a well-preorder (and so in particular a connected preorder) introduces a hidden "quantitative" feature; indeed, any such preorder can be equivalently described using a plausibility map as in e.g., [49], assigning ordinals to states. Our answer is that, first, the specific ordinals will not play any role in our definition of a dynamic belief update; and second, all our postulates can be given a justification in purely qualitative terms, using conditional beliefs. The transitivity condition for $\leq$ is just a *consistency* requirement imposed on a rational agent's conditional beliefs. And the existence of minimal elements in any non-empty subset is simply

the natural extension of the above setting to *general* conditional beliefs, not only conditions involving two states: more specifically, for any possible condition $P \subseteq S$ about a system S, the S-proposition $\mathrm{Min}_{\leq} P$ is simply a way to encode everything that the agent would believe about the current state of the system, if she was given the information that the state satisfied condition P.

Note on other models in the literature. Our models are the same as Board's "belief revision structures" [19], i.e., nothing but "Spohn models" as in [49], but with a purely relational description. Spohn models are usually described in terms of a map assigning ordinals to states. But giving such a map is equivalent to introducing a well pre-order $\leq$ on states, and it is easy to see that all the relevant information is captured by this order.

Our conditions on the preorder $\leq$ can also be seen as a *semantic analogue* of Grove's conditions for the (relational version of) his models in [34]. The standard formulation of Grove models is in terms of a "system of spheres" (weakening Lewis' similar notion), but it is equivalent (as proved in [34]) to a relational formulation. Grove's postulates are still *syntax-dependent*, e.g., existence of minimal elements is required only for subsets that are *definable* in his language: this is the so-called "smoothness" condition, which is weaker than our "well-preordered" condition. We prefer a purely semantic condition, independent of the choice of a language, both for reasons of elegance and simplicity and because we want to be able to consider more than one language for the same structure.[7] So, following [19, 52] and others, we adopt the natural semantic analogue of Grove's condition, simply requiring that *every* subset has minimal elements: this will allow our conditional operators to be well-defined on sentences of *any* extension of our logical language.

Note that the minimality condition implies, by itself, that the relation $\leq$ is both *reflexive* (i.e., $s \leq s$ for all $s \in S$) and *connected*[8] (i.e., either $s \leq t$ or $t \leq s$, for all $s, t \in S$). In fact, a "well-preorder" is the same as a *connected, transitive, well-founded*[9] *relation*, which is the setting proposed in [19] for a logic of conditional beliefs equivalent to our logic CDL below. Note also that, when the set S is *finite*, a well-preorder is nothing but a *connected preorder*. This shows that our notion of frame subsumes, not only Grove's setting, but also some of the other settings proposed for conditionalization.

[7] Imposing syntactic-dependent conditions in the very definition of a class of structures makes the definition meaningful only for one language; or else, the meaning of what, say, a plausibility model is won't be *robust*: it will change whenever one wants to extend the logic, by adding a few more operators. This is very undesirable, since then one cannot compare the expressivity of different logics on the same class of models.

[8] In the Economics literature, connectedness is called "completeness", see e.g., [19].

[9] I.e., there exists no infinite descending chain $s_0 > s_1 > \cdots$.

2.2 Multi-agent plausibility models

In the multi-agent case, *we cannot exclude from the model the states that are known to be impossible* by some agent a: they may still be considered possible by a second agent b. Moreover, they might still be relevant for a's beliefs/knowledge about what b believes or knows. So, in order to define an agent's knowledge, we cannot simply quantify over *all* states, as we did above: instead, we need to consider, as usually done in the Kripke-model semantics of knowledge, only the "possible" states, i.e., the ones that are *indistinguishable* from the real state, as far as a given agent is concerned. It is thus natural, in the multi-agent context, to explicitly specify the agents' epistemic indistinguishability relations $\sim_a$ (labeled with the agents' names) as part of the basic structure, in addition to the plausibility relations $\leq_a$. Taking this natural step, we obtain *epistemic plausibility frames* $(S, \sim_a, \leq_a)$. As in the case of a single agent, specifying epistemic relations turns out to be *superfluous*: the relations $\sim_a$ can be recovered from the relations $\leq_a$. Hence, we will simplify the above structures, obtaining the equivalent setting of *multi-agent plausibility frames* $(S, \leq_a)$.

Before going on to define these notions, observe that it doesn't make sense anymore to require the plausibility relations $\leq_a$ to be connected (and even less sense to require them to be well-preordered): if two states s, t are *distinguishable* by an agent a, i.e., $s \not\sim_a t$, then a will never consider both of them as epistemically possible in the same time. If she was given the information that the real state is either s or t, agent a will immediately *know* which of the two: if the real state was s, she would be able to distinguish this state from t, and would thus know the state was s; similarly, if the real state was t, she would know it to be t. Her beliefs will play no role in this, and it would be meaningless to ask her which of the two states is more plausible to her. So only the states in the same $\sim_a$-equivalence class could, and should, be $\leq_a$-comparable; i.e., $s \leq_a t$ *implies* $s \sim_a t$, and the restriction of $\leq_a$ to each $\sim_a$-equivalence class is connected. Extending the same argument to arbitrary conditional beliefs, we can see that *the restriction of* $\leq_a$ *to each* $\sim_a$*-equivalence class must be well-preordered.*

Epistemic plausibility frames. Let $\mathcal{A}$ be a finite set of labels, called *agents*. A *epistemic plausibility frame* over $\mathcal{A}$ (EPF, for short) is a structure $\mathbf{S} = (S, \sim_a, \leq_a)_{a \in \mathcal{A}}$, consisting of a set S of "states", endowed with a family of equivalence relations $\sim_a$, called *epistemic indistinguishability relations*, and a family of *plausibility relations* $\leq_a$, both labeled by "agents" and assumed to satisfy two conditions: (1) $\leq_a$-comparable states are $\sim_a$-indistinguishable (i.e., $s \leq_a t$ implies $s \sim_a t$); (2) the restriction of each plausibility relation $\leq_a$ to each $\sim_a$-equivalence class is a well-preorder. As before, we use the notation $\mathrm{Min}_{\leq_a} P$ for the set of $\leq_a$-minimal elements of P. We write $s <_a t$ iff $s \leq_a t$ but $t \not\leq_a s$ (the *"strict" plausibility rela-*

tion), and write $s \cong_a t$ iff both $s \leq_a t$ and $t \leq_a s$ (the *"equi-plausibility" relation*). The notion of *epistemic plausibility models* (EPM, for short) is defined in the same way as the plausibility models in the previous section.

Epistemic plausibility models. We define a *(multi-agent) epistemic plausibility model* (EPM, for short) as a multi-agent EPF together with a valuation over it (the same way that single-agent plausibility models were defined in the previous section).

It is easy to see that our definition of EPFs includes superfluous information: in an EPF, the knowledge relation $\sim_a$ can be recovered from the plausibility relation $\leq_a$, via the following rule:

$$s \sim_a t \text{ iff either } s \leq_a t \text{ or } t \leq_a s\,.$$

In other words, two states are indistinguishable for a iff they are *comparable* (with respect to $\leq_a$).

So, in fact, one could present epistemic plausibility frames simply as *multi-agent plausibility frames*. To give this alternative presentation, we use, for *any preorder relation* $\leq$, the notation $\sim$ for the associated *comparability relation*

$$\sim \;:= \leq \cup \geq$$

(where $\geq$ is the converse of $\leq$). A *comparability class* is a set of the form $\{t : s \leq t \text{ or } t \leq s\}$, for some state s. A relation $\leq$ is called *locally well-preordered* if it is a preorder such that its restriction to each comparability class is well-preordered. Note that, when the underlying set S is *finite*, a locally well-preordered relation is nothing but a *locally connected preorder*: a preorder whose restrictions to any comparability class are connected. More generally, *a locally well-preordered relation is the same as a locally connected and well-founded preorder.*

Multi-agent plausibility frames. A *multi-agent plausibility frame* (MPF, for short) is a structure $(S, \leq_a)_{a\in\mathcal{A}}$, consisting of a set of states S together with a family of locally well-preordered relations $\leq_a$, one for each agent $a \in \mathcal{A}$. Oliver Board [19] calls multi-agent plausibility frames "belief revision structures". A *multi-agent plausibility model* (MPM, for short) is an MPF together with a valuation map.

Bijective correspondence between EPFs and MPFs. *Every MPF can be canonically mapped into an EPF*, obtained by defining epistemic indistinguishability via the above rule ($\sim_a \;:= \leq_a \cup \geq_a$). Conversely, every EPF gives rise to an MPF, via the map that "forgets" the indistinguishability structure. It is easy to see that these two maps are the inverse of each other. Consequently, from now on we *identify MPFs and EPFs, and similarly identify MPMs and EPMs*; e.g., we can talk about "knowledge",

"(conditional) belief" etc. in an MPM, defined in terms of the associated EPM.

So from now on we identify the two classes of models, via the above canonical bijection, and talk about "plausibility models" in general. One can also see how this approach relates to another widely adopted definition for conditional beliefs; in [19, 24, 14], this definition involves the assumption of a *"local plausibility" relation at a given state* $s \leq_a^w t$, to be read as: "at state w, agent a considers state s at least as plausible as state t ". Given such a relation, the conditional belief operator is usually defined in terms that are equivalent to putting $s \rightarrow_a^P t$ iff $t \in \mathrm{Min}_{\leq_a^s} P$. One could easily restate our above definition in this form, by taking:

$$s \leq_a^w t \quad \text{iff} \quad \text{either } w \not\sim_a t \text{ or } s \leq_a t.$$

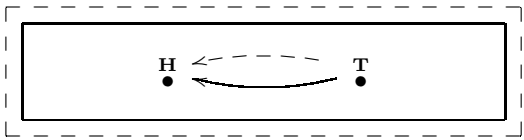
The converse problem is studied in [19], where it is shown that, if full introspection is assumed, then one can recover "uniform" plausibility relations $\leq_a$ from the relations $\leq_a^w$.

Information cell. The equivalence relation $\sim_a$ induces a partition of the state space S, called *agent a's information partition.* We denote by $s(a)$ the *information cell* of s in a's partition, i.e., the $\sim_a$-equivalence class of s:

$$s(a) := \{t \in S : s \sim_a t\}.$$

The information cell $s(a)$ captures *all the knowledge possessed by the agent* at state s: when the actual state of the system is s, then agent a knows only the state's equivalence class $s(a)$.

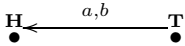
Example 2.1. Alice and Bob play a game, in which an anonymous referee puts a coin on the table, lying face up but in such a way that the face is covered (so Alice and Bob cannot see it). Based on previous experience, (it is common knowledge that) Alice and Bob believe that the upper face is Heads (since e.g., they noticed that the referee had a strong preference for Heads). And in fact, they're right: the coin lies Heads up. Neglecting the anonymous referee, the EPM for this example is the following model **S**:



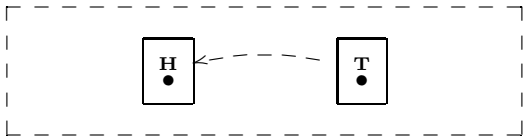
Here, the arrows represent *converse plausibility relations* $\geq$ *between distinct states only* (going from less plausible to more plausible states): since these are always reflexive, we choose to *skip all the loops* for convenience. The squares represent the *information cells* for the two agents. Instead of labels,

we use *dashed arrows and squares for Alice*, while using *continuous arrows and squares for Bob*. In this picture, the actual state of the system is the state s on the left (in which **H** is true). Henceforth, in our other examples, we will refer to this particular plausibility model as **S**.

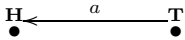
By deleting the squares, we obtain a representation of the corresponding MPM, also denoted by **S** (where we now use labels for agents instead of different types of lines):



Example 2.2. In front of Alice, the referee shows the face of the coin to Bob, but Alice cannot see the face. The EPM is now the following model **W**:



while the MPM is



Since Bob now knows the state of the coin, his local plausibility relation consists only of loops, and hence we have no arrows for Bob in this diagrammatic representation.

(Conditional) doxastic appearance and (conditional) doxastic accessibility. As in the previous section, we can define a doxastic and epistemic accessibility relations, except that now we have to select, for each state s, the most plausible states in its information cell $s(a)$ (instead of the most plausible in S). For this, it is convenient to introduce some notation and terminology: the *doxastic appearance* of state s to agent a is the set

$$s_a := \mathrm{Min}_{\leq_a} s(a)$$

of the "most plausible" states that are consistent with the agent's knowledge at state s. The doxastic appearance of s captures *the way state s appears to the agent*, or (in the language of Belief Revision) *the agent's current "theory" about the world s*. We can extend this to capture *conditional beliefs* (in full generality), by associating to each S-proposition $P \subseteq S$ and each state $s \in S$ the *conditional doxastic appearance* s_a^P *of state s to agent a, given (information)* P. This can be defined as the S-proposition

$$s_a^P := \mathrm{Min}_{\leq_a} s(a) \cap P$$

given by the set of all $\leq_a$-minimal states of $s(a) \cap P$: these are the "most plausible" states satisfying P that are consistent with the agent's knowledge at state s. The conditional appearance S_a^P gives *the agent's revised theory (after learning P) about the world s*. We can put these in a relational form, by defining *doxastic accessibility relations* $\rightarrow_a$, $\rightarrow_a^P$, as follows:

$$s \rightarrow_a t \text{ iff } t \in s_a\,,$$
$$s \rightarrow_a^P t \text{ iff } t \in s_a^P\,.$$

Knowledge and (conditional) belief. As before, we define the *knowledge* and *(conditional) belief* operators for an agent a as the Kripke modalities for a's epistemic and (conditional) doxastic accessibility relations:

$$K_a P := [\sim_a]P = \{s \in S : s(a) \subseteq P\}\,,$$
$$B_a P := [\rightarrow_a]P = \{s \in S : s_a \subseteq P\}\,,$$
$$B_a^Q P := [\rightarrow_a^Q]P = \{s \in S : s_a^Q \subseteq P\}.$$

We also need a notation for the *dual of the K modality* ("epistemic possibility"):

$$\tilde{K}_a P := \neg K_a \neg P.$$

Doxastic propositions. Until now, our notion of proposition is "local", being specific to a given model: we only have "S-propositions" for each model S. As long as the model is fixed, this notion is enough for interpreting sentences over the given model. But, since later we will proceed to study systematic *changes* of models (when dealing with *dynamic* belief revision), we need a notion of proposition that is not confined to one model, but makes sense on *all* models:

A *doxastic proposition* is a map $\mathbf{P}$ assigning to each plausibility model $\mathbf{S}$ some S-proposition $\mathbf{P_S} \subseteq S$. We write $s \models_{\mathbf{S}} \mathbf{P}$, and say that the proposition $\mathbf{P}$ is true at $s \in \mathbf{S}$, iff $s \in (\mathbf{P})_{\mathbf{S}}$. We skip the subscript and write $s \models \mathbf{P}$ when the model is understood.

We denote by Prop the family of all doxastic propositions. All the Boolean operations on S-propositions as sets can be *lifted* pointwise to operations on Prop: in particular, we have the "always true" $\top$ and "always false" $\bot$ propositions, given by $(\bot)_{\mathbf{S}} := \varnothing, (\top)_{\mathbf{S}} := S$, negation $(\neg\mathbf{P})_{\mathbf{S}} := S \setminus \mathbf{P_S}$, conjunction $(\mathbf{P} \wedge \mathbf{Q})_{\mathbf{S}} := \mathbf{P_S} \cap \mathbf{Q_S}$, disjunction $(\mathbf{P} \vee \mathbf{Q})_{\mathbf{S}} := \mathbf{P_S} \cup \mathbf{Q_S}$ and all the other standard Boolean operators, including *infinitary* conjunctions and disjunctions. Similarly, we can define pointwise the *epistemic and (conditional) doxastic modalities*: $(K_a\mathbf{P})_{\mathbf{S}} := K_a\mathbf{P_S}$, $(B_a\mathbf{P})_{\mathbf{S}} := B_a\mathbf{P_S}$, $(B_a^{\mathbf{Q}}\mathbf{P})_{\mathbf{S}} := B_a^{\mathbf{Q_S}}\mathbf{P_S}$. It is easy to check that we have: $B_a\mathbf{P} = B_a^{\top}\mathbf{P}$. Finally, the relation of *entailment* $\mathbf{P} \models \mathbf{Q}$ between doxastic propositions is given pointwise by inclusion: $\mathbf{P} \models \mathbf{Q}$ iff $\mathbf{P_S} \subseteq \mathbf{Q_S}$ for all $\mathbf{S}$.

2.3 Safe belief and the Defeasibility Theory of Knowledge

Ever since Plato's *identification of knowledge with "true justified (or justifiable) belief"* was shattered by Gettier's celebrated counterexamples [32], philosophers have been looking for the "missing ingredient" in the Platonic equation. Various authors identify this missing ingredient as "robustness" (Hintikka [35]), "indefeasibility" (Klein [38], Lehrer [40], Lehrer and Paxson [41], Stalnaker [52]) or "stability" (Rott [46]). According to this *defeasibility theory of knowledge* (or "stability theory", as formulated by Rott), a belief counts as "knowledge" if it is *stable under belief revision with any new evidence*: "if a person has knowledge, than that person's justification must be sufficiently strong that it is not capable of being defeated by evidence that he does not possess" (Pappas and Swain [43]).

One of the problems is interpreting what "evidence" means in this context. There are at least two natural interpretations, each giving us a concept of "knowledge". The first, and the most common[10], interpretation is to take it as meaning "any *true* information". The resulting notion of "knowledge" was formalized by Stalnaker in [52], and defined there as follows: "an agent knows that φ if and only if φ is true, she believes that φ, and she continues to believe φ if any *true* information is received". This concept differs from the usual notion of knowledge ("Aumann knowledge") in Computer Science and Economics, by the fact that it does not satisfy the laws of the modal system S5 (in fact, negative introspection fails); Stalnaker shows that the complete modal logic of this modality is the modal system S4.3. As we'll see, this notion ("Stalnaker knowledge") corresponds to what we call "safe belief" $\Box P$. On the other hand, another natural interpretation, considered by at least one author [46], takes "evidence" to mean "*any* proposition", i.e., to include possible *misinformation*: "real knowledge" should be robust even in the face of false evidence. As shown below, this corresponds to our "knowledge" modality KP, which could be called "absolutely unrevisable belief". This is a partition-based concept of knowledge, identifiable with "Aumann knowledge" and satisfying all the laws of S5. In other words, this last interpretation provides a perfectly decent "defeasibility" defense of S5 and of negative introspection!

In this paper, we adopt the pragmatic point of view of the formal logician: instead of debating which of the two types of "knowledge" is the real one, we simply formalize both notions in a common setting, compare them, axiomatize the logic obtained by combining them and use their joint strength to express interesting properties. Indeed, as shown below, conditional beliefs can be *defined* in terms of knowledge *only* if we combine both the above-mentioned types of "knowledge".

[10] This interpretation is the one virtually adopted by all the proponents of the defeasibility theory, from Lehrer to Stalnaker.

Knowledge as unrevisable belief. Observe that, for all propositions $\mathbf{P}$, we have

$$K_a\mathbf{Q} = \bigwedge_{\mathbf{P}} B_a^{\mathbf{P}}\mathbf{Q}$$

(where the conjunction ranges over *all* doxastic propositions), or equivalently, we have for every state s in every model $\mathbf{S}$:

$$s \models K_a\mathbf{Q} \quad \text{iff} \quad s \models B_a^{\mathbf{P}}\mathbf{Q} \text{ for all } \mathbf{P}. \tag{2.1}$$

This gives a characterization of *knowledge as "absolute" belief, invariant under any belief revision*: a given belief is "known" iff it cannot be revised, i.e., it would be still believed in any condition.[11] Observe that this resembles the defeasibility analysis of knowledge, but only if we adopt the *second interpretation* mentioned above (taking "evidence" to include misinformation). Thus, our "knowledge" is more robust than Stalnaker's: it resists any belief revision, not capable of being defeated by *any* evidence (including false evidence). This is a very "strong" notion of knowledge (implying "absolute certainty" and full introspection), which seems to us to fit better with the standard usage of the term in Computer Science literature. Also, unlike the one in [52], our notion of knowledge *is negatively introspective*.

Another identity[12] that can be easily checked is:

$$K_a\mathbf{Q} = B_a^{\neg\mathbf{Q}}\mathbf{Q} = B_a^{\neg\mathbf{Q}}\bot \tag{2.2}$$

(where $\bot$ is the "always false" proposition). This captures in a different way the "absolute un-revisability" of knowledge: something is "known" if it is believed even if conditionalizing our belief with its negation. In other words, this simply expresses the *impossibility* of accepting its negation as evidence (since such a revision would lead to an inconsistent belief).

Safe belief. To capture "Stalnaker knowledge", we introduce the Kripke modality $\Box_a$ associated to the converse $\geq_a$ of the plausibility relation, going from any state s to all the states that are "at least as plausible" as s. For S-propositions $P \subseteq S$ over any given model $\mathbf{S}$, we put

$$\Box_a P := [\geq_a]P = \{s \in S : t \in P \text{ for all } t \leq_a s\},$$

and this induces pointwise an operator $\Box_a\mathbf{P}$ on doxastic propositions. We read $s \models \Box_a\mathbf{P}$ as saying that: *at state s, agent a's belief in* $\mathbf{P}$ *is safe*; or at

[11] This of course assumes agents to be "rational" in a sense that excludes "fundamentalist" or "dogmatic" beliefs, i.e., beliefs in unknown propositions but refusing any revision, even when contradicted by facts. But this "rationality" assumption is already built in our plausibility models, which satisfy an epistemically friendly version of the standard AGM postulates of rational belief revision. See [8] for details.

[12] This identity corresponds to the definition of "necessity" in [50] in terms of doxastic conditionals.

state s, a safely believes that $\mathbf{P}$. We will explain this reading below, but first observe that: $\Box_a$ is an $S4$-modality (since $\geq_a$ is reflexive and transitive), but not necessarily $S5$; i.e., *safe beliefs are truthful* ($\Box_a\mathbf{P} \models \mathbf{P}$) and *positively introspective* ($\Box_a\mathbf{P} \models \Box_a\Box_a\mathbf{P}$), but not necessarily negatively introspective: in general, $\neg\Box_a\mathbf{P} \not\models \Box_a\neg\Box_a\mathbf{P}$.

Relations between knowledge, safe belief and conditional belief. First, *knowledge entails safe belief*

$$K_a\mathbf{P} \models \Box_a\mathbf{P}\,,$$

and *safe belief entails belief*

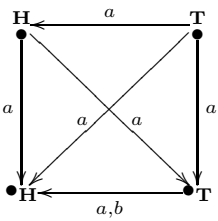
$$\Box_a\mathbf{P} \models B_a\mathbf{P}.$$

The last observation can be strengthened to characterize safe belief in a similar way to the above characterization (2.1) of knowledge (as belief invariant under any revision): *safe beliefs are precisely the beliefs which are persistent under revision with any true information.* Formally, this says that, for every state s in every model $\mathbf{S}$, we have

$$s \models \Box_a\mathbf{Q} \quad \text{iff:} \quad s \models B_a^{\mathbf{P}}\mathbf{Q} \text{ for every } \mathbf{P} \text{ such that } s \models \mathbf{P} \qquad (2.3)$$

We can thus see that *safe belief coincides indeed with Stalnaker's notion of "knowledge"*, given by the first interpretation ("evidence as true information") of the defeasibility theory. As mentioned above, we prefer to keep the name "knowledge" for the strong notion (which gives absolute certainty), and call this weaker notion "safe belief": indeed, these are beliefs that are "safe" to hold, in the sense that no future learning of truthful information will force us to revise them.

Example 2.3 (Dangerous Knowledge)**.** This starts with the situation in Example 2.1 (when none of the two agents has yet seen the face of the coin). Alice has to get out of the room for a minute, which creates an opportunity for Bob to quickly raise the cover in her absence and take a peek at the coin. He does that, and so he sees that the coin is Heads up. After Alice returns, she obviously doesn't know whether or not Bob took a peek at the coin, but she believes he didn't do it: taking a peek is against the rules of the game, and so she trusts Bob not to do that. The model is now rather complicated, so we only represent the MPM:



Let us call this model $\mathbf{S}'$. The actual state s'_1 is the one in the upper left corner, in which Bob took a peek and saw the coin Heads up, while the state t'_1 in the upper right corner represents the other possibility, in which Bob saw the coin lying Tails up. The two lower states s'_2 and t'_2 represent the case in which Bob *didn't take a peek*. Observe that the above drawing includes the (natural) assumption that Alice keeps her previous belief that the coin lies Heads up (since there is no reason for her to change her mind). Moreover, we assumed that she will keep this belief even if she'd be told that Bob took a peek: this is captured by the a-arrow from t'_1 to s'_1. This seems natural: Bob's taking a peek doesn't change the upper face of the coin, so it shouldn't affect Alice's prior belief about the coin.

In both Examples 2.1 and 2.3 above, Alice holds a *true belief* (at the real state) that the coin lies Heads up: the actual state satisfies $B_a\mathbf{H}$. In both cases, this true belief is *not knowledge* (since Alice doesn't know the upper face), but nevertheless in Example 2.1, this belief is *safe* (although it is *not known by the agent to be safe*): no additional truthful information (about the real state s) can force her to revise this belief. (To see this, observe that any *new* truthful information would reveal to Alice the real state s, thus confirming her belief that Heads is up.) So in the model $\mathbf{S}$ from Example 2.1, we have $s \models \Box_a\mathbf{H}$ (where s is the actual state). In contrast, in Example 2.2, Alice's belief (that the coin is Heads up), though true, is *not safe*. There is some piece of correct information (about the real state s'_1) which, if learned by Alice, would make her change this belief: we can represent this piece of correct information as the doxastic proposition $\mathbf{H} \to \mathbf{K_b H}$. It is easy to see that the actual state s'_1 of the model $\mathbf{S}'$ satisfies the proposition $B_a^{\mathbf{H}\to\mathbf{K_b H}}\mathbf{T}$ (since $(\mathbf{H} \to \mathbf{K_b H})_{\mathbf{S}'} = \{s'_1, t'_1, t'_2\}$ and the minimal state in the set $s'_1(a) \cap \{s'_1, s'_1, t'_2\} = \{s'_1, t'_1, t'_2\}$ is t'_2, which satisfies $\mathbf{T}$.) So, if given this information, Alice would come to wrongly believe that the coin is Tails up! This is an example of a *dangerous truth*: a true information whose learning can lead to wrong beliefs.

Observe that *an agent's belief can be safe without him necessarily knowing this* (in the "strong" sense of knowledge given by K): "safety" (similarly to "truth") is an *external* property of the agent's beliefs, that can be ascertained only by comparing his belief-revision system with reality. Indeed, *the only way* for an agent to *know a belief to be safe* is to actually *know it to be truthful*, i.e., to have actual *knowledge* (not just a belief) of its truth. This is captured by the valid identity

$$K_a\Box_a\mathbf{P} = K_a\mathbf{P}. \tag{2.4}$$

In other words: *knowing that something is safe to believe is the same as just knowing it to be true*. In fact, *all beliefs held by an agent "appear safe" to*

him: in order to believe them, he has to believe that they are safe. This is expressed by the valid identity

$$B_a \Box_a \mathbf{P} = B_a \mathbf{P} \tag{2.5}$$

saying that: *believing that something is safe to believe is the same as just believing it.* Contrast this with the situation concerning "knowledge": in our logic (as in most standard doxastic-epistemic logics), we have the identity

$$B_a K_a \mathbf{P} = K_a \mathbf{P}. \tag{2.6}$$

So *believing that something is known is the same as knowing it*!

The Puzzle of the Perfect Believer. The last identity is well-known and has been considered "paradoxical" by many authors. In fact, the so-called "Paradox of the Perfect Believer" in [33, 53, 36, 42, 54, 27] is based on it. For a "strong" notion of belief as the one we have here ("belief" = belief with certainty), it seems reasonable to assume the following "axiom":

$$B_a \varphi \rightarrow B_a K_a \varphi. \tag{?}$$

Putting this together with (2.6) above, we get a paradoxical conclusion:

$$B_a \varphi \rightarrow K_a \varphi. \tag{?!}$$

So this leads to a triviality result: *knowledge and belief collapse to the same thing*, and *all beliefs are always true*! One solution to the "paradox" is to reject (?), as an (intuitive but) *wrong* "axiom". In contrast, various authors [53, 36, 27, 54] accept (?) and propose other solutions, e.g., giving up the principle of "negative introspection" for knowledge.

Our solution to the paradox, as embodied in the contrasting identities (2.5) and (2.6), combines the advantages of both solutions above: the "axiom" (?) is *correct if we interpret "knowledge" as safe belief* $\Box_a$, since then (?) becomes equivalent to identity (2.5) above; but then *negative introspection fails for this interpretation*! On the other hand, if we interpret "knowledge" as our K_a-modality then negative introspection holds; but then *the above "axiom" (?) fails*, and on the contrary we have the identity (2.6).

So, in our view, *the paradox of the perfect believer arises from the conflation of two different notions of "knowledge":* "Aumann" (partition-based) knowledge and "Stalnaker" knowledge (i.e., safe belief).

(Conditional) beliefs in terms of "knowledge" notions. An important observation is that *one can characterize/define (conditional) beliefs only in terms of our two "knowledge" concepts (K and* $\Box$*)*: For simple beliefs, we have

$$B_a \mathbf{P} \;=\; \tilde{K}_a \Box_a \mathbf{P} \;=\; \Diamond_a \Box_a \mathbf{P}\,,$$

recalling that $\tilde{K}_a\mathbf{P} = \neg K_a \neg \mathbf{P}$ is the Diamond modality for K_a, and $\Diamond_a \mathbf{P} = \neg \Box_a \neg \mathbf{P}$ is the Diamond for $\Box_a$.

The equivalence $B_a\mathbf{P} = \Diamond_a \Box_a \mathbf{P}$ has recently been observed by Stalnaker in [52], who took it as the basis of a philosophical analysis of "belief" in terms of "defeasible knowledge" (i.e., safe belief). Unfortunately, this analysis does not apply to conditional belief: one can easily see that *conditional belief cannot be defined in terms of safe belief only*! However, one *can generalize* the identity $B_a\mathbf{P} = \tilde{K}_a \Box_a \mathbf{P}$ above, defining conditional belief in terms of *both our "knowledge" concepts*:

$$B_a^{\mathbf{P}}\mathbf{Q} \;=\; \tilde{K}_a\mathbf{P} \to \tilde{K}_a(\mathbf{P} \wedge \Box_a(\mathbf{P} \to \mathbf{Q}))\,. \tag{2.7}$$

2.4 Other modalities and doxastic attitudes

From a modal logic perspective, it is natural to introduce the Kripke modalities $[>_a]$ and $[\cong_a]$ for the other important relations (strict plausibility and equiplausibility): For S-propositions $P \subseteq S$ over a given model $\mathbf{S}$, we put

$$[>_a]P := \{s \in S : t \in P \text{ for all } t <_a s\},$$
$$[\cong_a]P := \{s \in S : t \in P \text{ for all } t \cong_a s\},$$

and as before these pointwise induce corresponding operators on Prop. The intuitive meaning of these operators is not very clear, but they can be used to define other interesting modalities, capturing various "doxastic attitudes".

Weakly safe belief. We can define a *weakly safe belief* operator $\Box^{\text{weak}}\mathbf{P}$ in terms of the strict order by putting:

$$\Box_a^{\text{weak}}\mathbf{P} \;=\; \mathbf{P} \wedge [>_a]\mathbf{P}\,.$$

Clearly, this gives us the following truth clause:

$$s \models \Box_a^{\text{weak}}\ \mathbf{P} \text{ iff:} \quad s \models \mathbf{P} \text{ and } t \models \mathbf{P} \text{ for all } t < s\,.$$

But a more useful characterization is the following:

$$s \models \Box_a^{\text{weak}}\mathbf{Q} \quad \text{iff:} \quad s \models \neg B_a^{\mathbf{P}} \neg \mathbf{Q} \text{ for every } \mathbf{P} \text{ such that } s \models \mathbf{P}\,.$$

So "weakly safe beliefs" are *beliefs which (might be lost but) are never reversed (into believing the opposite) when revising with any true information.*

The unary revision operator. Using the strict plausibility modality, we can also define a unary "belief revision" modality $*_a$, which in some sense *internalizes the standard (binary) belief revision operator*, by putting:

$$*_a\mathbf{P} \;=\; \mathbf{P} \wedge [>_a]\neg\mathbf{P}\,.$$

This gives us the following truth clause:

$$s \models *_a\mathbf{P} \text{ iff } s \in s_a^{\mathbf{P}} .$$

It is easy to see that $*_a\mathbf{P}$ *selects from any given information cell* $s(a)$ *precisely those states that satisfy agent a's revised theory* s_a^P:

$$*_aP \cap s(a) = s_a^P .$$

Recall that $s_a^P = \mathrm{Min}_{\leq_a} s(a) \cap P$ is the conditional appearance of s to a given P, representing the agent's "revised theory" (after revision with P) about s. This explains our interpretation: the proposition $*_a\mathbf{P}$ is a *complete description of the agent's P-revised "theory" about the current state.*

Another interesting identity is the following:

$$B_a^{\mathbf{P}}\mathbf{Q} \;=\; K_a(*_a\mathbf{P} \rightarrow \mathbf{Q}) . \tag{2.8}$$

In other words: $\mathbf{Q}$ is *a conditional belief (given a condition* $\mathbf{P}$*) iff it is a known consequence of the agent's revised theory (after revision with* $\mathbf{P}$*).*

Degrees of belief. Spohn's "degrees of belief" [49] were captured by Aucher [2] and van Ditmarsch [24] using logical operators $B_a^n\mathbf{P}$. Intuitively, 0-belief $B_a^0\mathbf{P}$ is the same as simple belief $B_a\mathbf{P}$; 1-belief $B_a^1\mathbf{P}$ means that $\mathbf{P}$ is believed conditional on learning that not all the 0-beliefs are true etc. Formally, this can be introduced e.g., by defining by induction a sequence of appearance maps s_a^n for all states s and natural numbers n:

$$s_a^0 = \mathrm{Min}_{\leq_a} s(a) \ , \ s_a^n = \mathrm{Min}_{\leq_a}\Big(s(a) \setminus \bigcup_{i<n} s_a^i\Big)$$

and defining

$$s \models B_a^n\mathbf{P} \text{ iff } t \models \mathbf{P} \text{ for all } t \in s_a^n.$$

A state s has degree of belief n if we have $s \in s_a^n$. An interesting observation is that the *finite degrees of belief* $B_a^n\mathbf{P}$ *can be defined using the unary revision operator* $*_a\mathbf{P}$ *and the knowledge operator* K_a (and, as a consequence, they can be defined using the plausibility operator $[>_a]\mathbf{P}$ and the knowledge operator). To do this, first put inductively:

$$b_a^0 := *_a\top \ , \ b_a^n := *_a\Big(\bigwedge_{m<n} \neg b_a^m\Big) \text{ for all } n \geq 1$$

and then put

$$B_a^n\mathbf{P} \;:=\; \bigwedge_{m<n} \neg K_a(b_a^m \rightarrow \mathbf{P}) \wedge K_a(b_a^n \rightarrow \mathbf{P}) .$$

"Strong belief". Another important doxastic attitude can be defined in terms of knowledge and safe belief as:

$$Sb_a\mathbf{P} \;=\; B_a\mathbf{P} \wedge K_a(\mathbf{P} \rightarrow \Box_a\mathbf{P})\,.$$

In terms of the plausibility order, it means that *all the* $\mathbf{P}$*-states in the information cell* $s(a)$ *of* s *are bellow (more plausible than) all the non-*$\mathbf{P}$ *states in* $s(a)$ (and that, moreover, *there are such* $\mathbf{P}$*-states in* $s(a)$). This notion is called "strong belief" by Battigalli and Siniscalchi [13], while Stalnaker [51] calls it "robust belief". Another characterization of strong belief is the following

$s \models Sb_a\mathbf{Q}$ iff:

$s \models B_a\mathbf{Q}$ and $s \models B_a^{\mathbf{P}}\mathbf{Q}$ for every $\mathbf{P}$ such that $s \models \neg K_a(\mathbf{P} \rightarrow \neg\mathbf{Q})$.

In other words: *something is strong belief if it is believed and if this belief can only be defeated by evidence (truthful or not) that is known to contradict it.* An example is the "presumption of innocence" in a trial: requiring the members of the jury to hold the accused as "innocent until proven guilty" means asking them to start the trial with a "strong belief" in innocence.

2.5 The logic of conditional beliefs

The logic CDL ("conditional doxastic logic") introduced in [8] is a logic of conditional beliefs, equivalent to the strongest logic considered in [19]. The *syntax* of CDL (without common knowledge and common belief operators[13]) is:

$$\varphi \;:=\; p \mid \neg\varphi \mid \varphi \wedge \varphi \mid B_a^{\varphi}\varphi$$

while the *semantics* is given by an *interpretation map* associating to each sentence φ of CDL a doxastic proposition $\|\varphi\|$. The definition is by induction, in terms of the obvious compositional clauses (using the doxastic operators $B_a^{\mathbf{P}}\mathbf{Q}$ defined above).

In this logic, *knowledge and simple (unconditional) belief are derived operators*, defined as abbreviations by putting $K_a\varphi := B_a^{\neg\varphi}\varphi$, $B_a\varphi := B_a^{\top}\varphi$ (where $\top := \neg(p \wedge \neg p)$ is some tautological sentence).

Proof system. In addition to the rules and axioms of propositional logic, the *proof system* of CDL includes the following:

Necessitation Rule:	From $\vdash \varphi$ infer $\vdash B_a^{\psi}\varphi$.
Normality:	$\vdash B_a^{\vartheta}(\varphi \rightarrow \psi) \rightarrow (B_a^{\vartheta}\varphi \rightarrow B_a^{\vartheta}\psi)$

[13] The logic in [8] has these operators, but for simplicity we decided to leave them aside in this presentation.

Truthfulness of Knowledge:	$\vdash K_a\varphi \to \varphi$
Persistence of Knowledge:	$\vdash K_a\varphi \to B_a^\vartheta\varphi$
Full Introspection:	$\vdash B_a^\vartheta\varphi \to K_aB_a^\vartheta\varphi$,
	$\vdash \neg B_a^\vartheta\varphi \to K_a\neg B_a^\vartheta\varphi$
Success of Belief Revision:	$\vdash B_a^\varphi\varphi$
Minimality of Revision:	$\vdash \neg B_a^\varphi\neg\psi \to (B_a^{\varphi\wedge\psi}\vartheta \leftrightarrow B_a^\varphi(\psi\to\vartheta))$

Proposition 2.4 (Completeness and Decidability)**.** The above system is complete for MPMs (and so also for EPMs). Moreover, it is decidable and has the finite model property.

Proof. The proof is essentially the same as in [19]. It is easy to see that the proof system above is equivalent to Board's strongest logic in [19] (the one that includes axiom for full introspection), and that our models are equivalent to the "full introspective" version of the semantics in [19]. Q.E.D.

2.6 The logic of knowledge and safe belief

The problem of finding a complete axiomatization of the logic of "defeasible knowledge" (safe belief) and conditional belief was posed as an *open question* in [19]. We answer this question here, by extending the logic CDL above to a complete logic $K\Box$ of *knowledge and safe belief*. Since this logic can *define* conditional belief, it is in fact equivalent to the logic whose axiomatization was required in [19]. Solving the question posed there becomes in fact trivial, once we observe that the higher-order definition of "defeasible knowledge" in [52, 19] (corresponding to our identity (1.3) above) is in fact equivalent to our simpler, first-order definition of "safe belief" as a Kripke modality.

Syntax and semantics. The *syntax* of the logic $K\Box$ is:

$$\varphi \;:=\; p \mid \neg\varphi \mid \varphi\wedge\varphi \mid \Box_a\varphi \mid K_a\varphi$$

while the *semantics* over plausibility models is given as for CDL, by inductively defining an interpretation map from sentences to doxastic propositions, using the obvious compositional clauses. *Belief and conditional belief are derived operators here, defined as abbreviations*:

$$\begin{aligned} B_a^\varphi\psi &:= \tilde{K}_a\varphi \to \tilde{K}_a(\varphi\wedge\Box_a(\varphi\to\psi))\,,\\ B_a\varphi &:= B_a^\top\varphi\,, \end{aligned}$$

where $\tilde{K}_a\varphi := \neg K_a\neg\varphi$ is the Diamond modality for K, and $\top = \neg(p\wedge\neg p)$ is some tautological sentence. So *the logic $K\Box$ is more expressive than CDL.*

Proof system. In addition to the rules and axioms of propositional logic, the *proof system* for the logic $K\Box$ includes the following:

- the Necessitation Rules for both K_a and $\Box_a$;
- the $S5$-axioms for K_a;
- the $S4$-axioms for $\Box_a$;
- $K_aP \to \Box_aP$;
- $K_a(P \vee \Box_aQ) \wedge K_a(Q \vee \Box_aP) \to K_aP \vee K_aQ$.

Theorem 2.5 (Completeness and Decidability)**.** The logic $K\Box$ is *(weakly) complete* with respect to MPMs (and so also with respect to EPMs). Moreover, it is *decidable* and has *the finite model property.*

Proof. A *non-standard frame (model)* is a structure $(S, \geq_a, \sim_a)_a$ (together with a valuation, in the case of models) such that $\sim_a$ are equivalence relations, $\geq_a$ are preorders, $\geq_a \subseteq \sim_a$ and the restriction of $\geq_a$ to each $\sim_a$-equivalence class is connected. For a logic with two modalities, $\Box_a$ for $\geq_a$ and K_a for the relation $\sim_a$, we can use well-known results in Modal Correspondence Theory to see that each of these semantic conditions corresponds to one of our modal axioms above. By general classical results on canonicity and modal correspondence[14], we immediately obtain *completeness for non-standard models. Finite model property for these non-standard models* follows from the same general results. But every *finite* strict preorder relation $>$ is well-founded, and an MPM is nothing but a non-standard model whose strict preorders $>_a$ are well-founded. So *completeness for ("standard") MPMs* immediately follows. Then we can use Proposition 2.4 above to obtain *completeness for EPMs.* Finally, *decidability* follows, in the usual way, from finite model property together with *completeness* (with respect to a *finitary* proof system) and with the *decidability of model-checking on finite models.* (This last property is obvious, given the semantics.) Q.E.D.

3 "Dynamic" Belief Revision

The revision captured by conditional beliefs is of a *static*, purely *hypothetical*, nature. We *cannot* interpret B_a^{φ} as referring to the agent's revised beliefs about the situation *after revision*; if we did, then the "Success" axiom

$$\vdash B_a^{\varphi}\varphi$$

would *fail for higher-level beliefs.* To see this, consider a "Moore sentence"

$$\varphi := p \wedge \neg B_a p\,,$$

[14] See e.g., [18] for the general theory of modal correspondence and canonicity.

saying that some fact p holds but that agent a doesn't believe it. The sentence φ is consistent, so it may very well happen to be true. But agent a's beliefs about the situation after learning that φ was true *cannot* possibly include the sentence φ itself: after learning this sentence, agent *a knows* p, and so he believes p, contrary to what φ asserts. Thus, after learning φ, agent *a knows that* φ *is false now* (after the learning). This directly contradicts the Success axiom: far from believing the sentence after learning it to be true, the agent (knows, and so he correctly) believes that it has become false. There is nothing paradoxical about this: sentences may obviously change their truth values, due to our actions. Since learning the truth of a sentence is itself an action, it is perfectly consistent to have a case in which learning changes the truth value of the very sentence that is being learnt. Indeed, this is always the case with Moore sentences. Though not paradoxical, the existence of Moore sentences shows that the "Success" axiom does not correctly describe a rational agent's (higher-level) beliefs about what is the case after a new truth is being learnt.

The only way to understand the "Success" axiom in the context of higher-level beliefs is to insist on the above-mentioned "static" interpretation of conditional belief operators B_a^{φ}, as expressing the agent's *revised belief* about how the state of the world *was before the revision*.

In contrast, a *belief update* is a dynamic form of belief revision, meant to capture the actual change of beliefs induced by learning: the updated belief is about the state of the world as it is *after the update*. As noticed in [29, 6, 5], the original model does not usually include enough states to capture all the epistemic possibilities that arise in this way. While in the previous section the models were kept unchanged during the revision, all the possibilities being already there (so that both the unconditional and the conditional beliefs *referred to the same model*), we now have to allow for belief updates that *change the original model*.

In [5], it was argued that *epistemic events should be modeled in essentially the same way as epistemic states*, and this common setting was taken to be given by *epistemic Kripke models*. Since in this paper we enriched our state models with doxastic plausibility relations to deal with (conditional) beliefs, it is natural to follow [5] into extending the similarity between actions and states to this setting, thus obtaining *(epistemic) action plausibility models*. The idea of such an extension was first developed in [2] (for a different notion of plausibility model and a different notion of update product), then generalized in [24], where many types of action plausibility models and notions of update product, that extend the so-called *Baltag-Moss-Solecki (BMS) update product* from [6, 5], are explored. But both these works are based on a *quantitative* interpretation of plausibility ordinals (as "degrees of belief"), and thus they define the various types of products using complex

formulas of transfinite ordinal arithmetic, for which no intuitive justification is provided.

In contrast, our notion of update product is a *purely qualitative one*, based on a *simple and intuitive relational definition*: the simplest way to define a total pre-order on a Cartesian product, given total pre-orders on each of the components, is to use either the *lexicographic* or the *anti-lexicographic* order. We choose the second option, as the closest in spirit to the classical AGM theory: it gives *priority to the new, incoming information* (i.e., to "actions" in our sense).[15] We justify this choice by interpreting the action plausibility model as representing the agent's *"incoming" belief*, i.e., the *belief-updating event*, which *"performs" the update, by "acting" on the "prior" beliefs* (as given in the state plausibility model).

3.1 Action models

An *action plausibility model*[16] (APM, for short) is a plausibility frame $(\Sigma, \leq_a)_{a\in\mathcal{A}}$ together with a *precondition map* $\mathrm{pre} : \Sigma \to \mathsf{Prop}$, associating to each element of Σ some doxastic proposition pre_σ. We call the elements of Σ *(basic) doxastic actions* (or "events"), and we call pre_σ the *precondition* of action σ. The basic actions $\sigma \in \Sigma$ are taken to represent *deterministic belief-revising actions* of a particularly simple nature. Intuitively, the precondition defines the *domain of applicability* of action σ: it can be executed on a state s iff s satisfies its precondition. The relations $\leq_a$ give the agents' beliefs about which actions are more plausible than others.

To model *non-determinism*, we introduce the notion of epistemic program. A *doxastic program over a given action model* $\mathbf{\Sigma}$ (or $\mathbf{\Sigma}$*-program*, for short) is simply a *set* $\Gamma \subseteq \Sigma$ of doxastic actions. We can think of doxastic programs as non-deterministic actions: each of the basic actions $\gamma \in \Gamma$ is a possible "deterministic resolution" of Γ. For simplicity, when $\Gamma = \{\gamma\}$ is a singleton, we ambiguously identify the program Γ with the action γ.

Observe that $\mathbf{\Sigma}$-programs $\Gamma \subseteq \Sigma$ are formally the "dynamic analogues" of $\mathbf{S}$-propositions $P \subseteq S$. So the dynamic analogue of the conditional doxastic appearance s_a^P (representing agent a's revised theory about state s, after revision with proposition P) is the set σ_a^Γ.

Interpretation: beliefs about changes encode changes of beliefs. The name "doxastic actions" might be a bit misleading, and from a philosophical perspective Johan van Benthem's term "doxastic events" seems more appropriate. The elements of a plausibility model do not carry information about agency or intentionality and cannot represent "real" actions in all their complexity, but only the *doxastic changes* induced by these actions: each of the nodes of the graph represents a *specific kind of change*

[15] This choice can be seen as a generalization of the so-called *"maximal-Spohn" revision*.
[16] Van Benthem calls this an "event model".

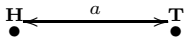
of beliefs (of all the agents). As in [5], we only deal here with pure "belief changes", i.e., actions that do not change the "ontic" facts of the world, but only the agents' beliefs.[17] Moreover, we think of these as *deterministic* changes: there is at most one output of applying an action to a state.[18] Intuitively, the precondition defines the *domain of applicability* of σ: this action can be executed on a state s iff s satisfies its precondition. The plausibility pre-orderings $\leq_a$ give *the agents' conditional beliefs about the current action.* But this should be interpreted as *beliefs about changes*, that *encode changes of beliefs.* In this sense, we use such "beliefs about actions" as a way to represent doxastic changes: the information about how the agent changes her beliefs is captured by our action plausibility relations. So we read $\sigma <_a \sigma'$ as saying that: if agent a is informed that either σ or σ' is currently happening, then she cannot distinguish between the two, but she believes that σ is in fact happening. As already mentioned, doxastic programs $\Gamma \subseteq \Sigma$ represent *non-deterministic* changes of belief. Finally, for an action σ and a program Γ, the program σ_a^Γ represents *the agent's revised theory (belief) about the current action σ after "learning" that (one of the deterministic resolutions γ in) Γ is currently happening.*

Example 3.1 (Private "Fair-Game" Announcements)**.** Let us consider the *action* that produced the situation represented in Example 2.2 above. In front of Alice, Bob looked at the coin, in such a way that (it was common knowledge that) only he saw the face. In the DEL literature, this is sometimes known as a "fair game" announcement: everybody is commonly aware that an insider (or a group of insiders) privately learns some information. It is "fair" since the outsiders are *not "deceived" in any way*: e.g., in our example, Alice knows that Bob looks at the coin (and he knows that she knows etc.). In other words, Bob's looking at the coin is not an "illegal" action, but one that obeys the (commonly agreed) "rules of the game". To make this precise, let us assume that this is happening in such a way that Alice has no strong beliefs about which of the two possible actions (Bob-seeing-Heads-up and Bob-seeing-Tails-up) is actually happening. Of course, we assumed that before this, she already believed that the coin lies Heads up, but apart from this we now assume that *the way the action (of "Bob looking") is happening gives her no indication of what face he is seeing.* We represent these actions using a two-node plausibility model Σ_2 (where as in the case of state models we draw arrows for the converse plausibility

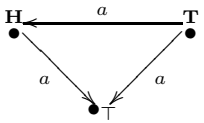
[17] We stress this is a minor restriction, and it is very easy to extend this setting to "ontic" actions. The only reason we stick with this restriction is that it simplifies the definitions, and that it is general enough to apply to all the actions we are interested here, and in particular to all *communication actions.*

[18] As in [5], we will be able to represent non-deterministic actions as sums (unions) of deterministic ones.

relations $\geq_a$, disregarding all the loops):



Example 3.2 (Fully Private Announcements)**.** Let us consider the *action* that produced the situation represented in Example 2.3 above. This was the action of Bob taking a peek at the coin, while Alice was away. Recall that we assumed that Alice *believed that nothing was really happening* in her absence (since she assumed Bob was playing by the rules), though obviously she *didn't know* this (that nothing was happening). In the DEL literature, this action is usually called *a fully private announcement*: Bob learns which face is up, while the outsider Alice believes nothing of the kind is happening. To represent this, we consider an action model $\mathbf{\Sigma}_3$ consisting of three "actions": the actual action σ in which Bob takes a peek and sees the coin lying Heads up; the alternative possible action ρ is the one in which Bob sees the coin lying Tails up; finally, the action τ is the one in which "nothing is really happening" (as Alice believes). The plausibility model $\mathbf{\Sigma}_3$ for this action is:



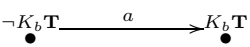
Here, the action σ is the one in the upper left corner, having precondition **H**: indeed, this can happen iff the coin is really lying Heads up; similarly, the action ρ in the upper right corner has precondition **T**, since it can only happen iff the coin is Tails up. Finally, the action τ is the lower one, having as precondition the "universally true" proposition $\top$: indeed, this action can always happen (since in it, nothing is really happening!). The plausibility relations reflect the agents' beliefs: in each case, both Bob and Charles know exactly what is happening, so their local plausibility relations are the identity (and thus we draw no arrows for them). Alice believes nothing is happening, so τ is the most plausible action for her (to which all her arrows are pointing); so she keeps her belief that **H** is the case, thus considering σ as more plausible than ρ.

Examples of doxastic programs. Consider the program $\Gamma = \{\sigma, \rho\} \subseteq \Sigma_3$ over the action model Σ_3 from Example 3.2. The program Γ represents *the action of "Bob taking a peek at the coin", without any specification of which face he is seeing.* Although expressed in a non-deterministic manner (as a collection of two possible actions, σ and ρ), this program corresponds is in fact *deterministic*, since in each possible state only one of the actions σ or ρ can happen: there is no state satisfying both **H** and **T**. The whole set Σ gives another doxastic program, one that is really non-deterministic:

it represents the non-deterministic choice of Bob between taking a peek and not taking it.

Appearance of actions and their revision: Examples. As an example of an agent's "theory" about an action, consider the appearance of action ρ to Alice: $\rho_a = \{\tau\}$. Indeed, if ρ happens (Bob taking a peek and sees the coin is Tails up), Alice believes that τ (i.e., nothing) is happening: this is the "apparent action", as far as Alice is concerned. As an example of a "revised theory" about an action, consider the conditional appearance ρ_a^Γ of ρ to Alice given the program $\Gamma = \{\sigma, \rho\}$ introduced above. It is easy to see that we have $\rho_a^\Gamma = \{\sigma\}$. This captures our intuitions about Alice's revised theory: if, while ρ was happening, she were told that Bob took a peek (i.e., she'd revise with Γ), then she would believe that he saw the coin lying Heads up (i.e., that σ happened).

Example 3.3 (Successful Lying)**.** Suppose now that, *after* the previous action, i.e., after we arrived in the situation described in Example 2.3, Bob sneakily announces: "I took a peek and saw the coin was lying *Tails up*". We formalize the content of this announcement as $K_b\mathbf{T}$, i.e., saying that "Bob knows the coin is lying Tails up". This is a *public announcement*, but *not a truthful one* (though it does convey some truthful information): it is a *lie*! We assume it is in fact a *successful lie*: it is common knowledge that, even after Bob admitted having taken a peek, Alice still believes him. This action is given by the *left node* in the following model Σ_4:



3.2 The action-priority update

We are ready to define our *update operation*, representing the way an action from a (action) plausibility model $\mathbf{\Sigma} = (\Sigma, \leq_a, \text{pre})_{a\in\mathcal{A}}$ "acts" on an input-state from a given (state) plausibility model $\mathbf{S} = (S, \leq_a, \|\cdot\|)_{a\in\mathcal{A}}$. We denote the updated state model by $\mathbf{S}\otimes\mathbf{\Sigma}$, and call it the *update product* of the two models. The construction is similar to a point to the one in [6, 5], and thus also somewhat similar to the ones in [2, 24]. In fact, the set of updated states, the updated valuation and the updated indistinguishability relation are *the same* in these constructions. The main difference lies in our definition of the *updated plausibility relation*, via the *Action Priority Rule*.

3.2.1 Updating single-agent models: the anti-lexicographic order

To warm up, let us first define the update product for the single-agent case. Let $\mathbf{S} = (S, \leq, \|\cdot\|)$ be a single-agent plausibility state model and let $\mathbf{\Sigma} = (\Sigma, \leq, \text{pre})$ be a single-agent plausibility action model.

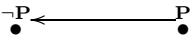
We represent the *states of the updated model* $\mathbf{S} \otimes \mathbf{\Sigma}$ as pairs (s, σ) of input-states and actions, i.e., as elements of the Cartesian product $S \times \Sigma$. This reflects that the basic actions in our action models are assumed to be *deterministic*: For a given input-state and a given action, there can only be at most one output-state. More specifically, we select the pairs which are *consistent, in the sense that the input-state satisfies the precondition of the action*. This is natural: the precondition of an action is a specification of its domain of applicability. So the *set of states* of $\mathbf{S} \otimes \mathbf{\Sigma}$ is taken to be

$$S \otimes \Sigma := \{(s, \sigma) : s \models_{\mathbf{S}} \mathrm{pre}(\sigma)\}.$$

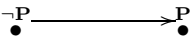
The *updated valuation* is essentially given by the *original valuation* from the input-state model: For all $(s, \sigma) \in S \otimes \Sigma$, we put $(s, \sigma) \models p$ iff $s \models p$. This "conservative" way to update the valuation expresses the fact that we only consider here actions that are *"purely doxastic"*, i.e., pure "belief changes", that do not affect the ontic "facts" of the world (captured here by atomic sentences).

We still need to define the updated plausibility relation. To motivate our definition, we first consider two examples:

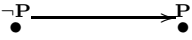
Example 3.4 (A Sample Case)**.** Suppose that we have two states $s, s' \in \mathbf{S}$ such that $s < s'$, $s \models \neg\mathbf{P}$, $s' \models \mathbf{P}$. This means that, if given the supplementary information that the real state is either s or s', the agent believes $\neg\mathbf{P}$:



Suppose then an event happens, in whose model there are two actions σ, σ' such that $\sigma > \sigma'$, $\mathrm{pre}_{\sigma} = \neg\mathbf{P}$, $\mathrm{pre}_{\sigma'} = \mathbf{P}$. In other words, if given the information that either σ or σ' is happening, the agent believes that σ' is happening, i.e., she believes that $\mathbf{P}$ is learnt. This part of the model behaves just like a *soft public announcement* of $\mathbf{P}$:

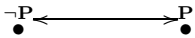


Naturally, we expect the agent to *change her belief* accordingly, i.e., her updated plausibility relation on states should now go the other way:

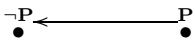


Example 3.5 (A Second Sample Case)**.** Suppose the initial situation was the same as above, but now the two actions σ, σ' are assumed to be equi-plausible: $\sigma \cong \sigma'$. This is a *completely unreliable announcement of* $\mathbf{P}$, in

which the veracity and the falsity of the announcement are equally plausible alternatives:



In the AGM paradigm, it is natural to expect the agents to *keep their original beliefs* unchanged after this event:



The anti-lexicographic order. Putting the above two sample cases together, we conclude that the updated plausibility relation should be the *anti-lexicographic preorder relation* induced on pairs $(s, \sigma) \in S \times \Sigma$ by the preorders on $\mathbf{S}$ and on $\mathbf{\Sigma}$, i.e.:

$$(s, \sigma) \leq (s', \sigma') \quad \text{iff:} \quad \text{either} \ \ \sigma < \sigma', \text{ or else } \sigma \cong \sigma' \text{ and } s \leq s'.$$

In other words, the updated plausibility order gives *"priority" to the action plausibility relation, and apart from this it keeps as much as possible the old order*. This reflects our commitment to an AGM-type of revision, in which the new information has priority over old beliefs. The "actions" represent here the "new information", although (unlike in AGM) this information comes in *dynamic form* (as action plausibility order), and so it is not fully reducible to its propositional content (the action's precondition). In fact, this is a generalization of one of the belief-revision policies encountered in the literature (the so-called "*maximal-Spohn revision*"). But, in the context of our qualitative (conditional) interpretation of plausibility models, we will argue below that this is essentially the only reasonable option.

3.2.2 Updating multi-agent models: the general case

In the multi-agent case, the construction of *the updated state space and updated valuation is the same as above*. But for the updated plausibility relation we need to take into account *a third possibility*: the case when either the initial states or the actions are *distinguishable*, belonging to *different information cells*.

Example 3.6 (A Third Sample Case)**.** Suppose that we have two states $s, s' \in \mathbf{S}$ such that $s \models \neg\mathbf{P}$, $s' \models \mathbf{P}$, but $s \not\sim_a s'$ are *distinguishable* (i.e., non-comparable):

This means that, if given the supplementary information that the real state is either s or s', the agent immediately *knows* which of the two is the real states, and thus *she knows whether* $\mathbf{P}$ *holds or not*. It is obvious

that, after any of the actions considered in the previous two examples, a perfect-recall agent *will continue to know* whether **P** held or not, and so *the output-states after σ and σ' will still be distinguishable (non-comparable).*

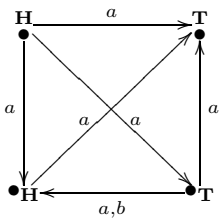
The "Action-Priority" Rule. Putting this together with the other sample cases, we obtain our update rule, in full generality:

$(s, \sigma) \leq_a (s', \sigma')$ iff either $\sigma <_a \sigma'$ and $s \sim_a s'$, or else $\sigma \cong_a \sigma'$ and $s \leq_a s'$

We regard this construction as the most natural analogue in a belief-revision context of the similar notion in [5, 6]. Following a suggestion of Johan van Benthem, we call this the Action-Priority Update Rule.

Sanity check: Examples 2.2 and 2.3 revisited. To check the correctness of our update operation, take first the update product $\mathbf{S} \otimes \mathbf{\Sigma}_2$ of the model **S** in Example 2.1 from the previous section with the action model $\mathbf{\Sigma}_2$ in Example 3.1 from the previous section. As predicted, the resulting state model is isomorphic to the model **W** from Example 2.2. Similarly, if $\mathbf{\Sigma}_3$ is the action model from Example 3.2, then we can see that the product $\mathbf{S} \otimes \mathbf{\Sigma}_3$ is isomorphic to the state model $\mathbf{S}'$ from Example 2.3.

"In-sanity check": Successful lying. Applying the action model $\mathbf{\Sigma}_4$ in Example 3.3, representing the "successful lying" action, to the state model $\mathbf{S}'$ from Example 2.3, we obtain indeed the intuitively correct output of "successful lying", namely the following model $\mathbf{S}' \otimes \mathbf{\Sigma}_4$:



Interpretation. As its name makes explicit, the Action-Priority Rule gives "priority" to the *action* plausibility relation. This is not an arbitrary choice, but it is motivated by our specific interpretation of action models, as embodied in our Motto above: *beliefs about changes* (i.e., the action plausibility relations) *are nothing but ways to encode changes of belief* (i.e., reversals of the original plausibility order). So *the (strict) order on actions encodes changes of order on states.* The Action-Priority Rule is a consequence of this interpretation: it just says that a strong plausibility order $\sigma <_a \sigma'$ on actions corresponds indeed to a change of ordering, (from whatever the ordering was) between the original (indistinguishable) input-states $s \sim_a s'$,

to the order $(s,\sigma) <_a (s',\sigma')$ between output-states; while equally plausible actions $\sigma \cong_a \sigma'$ will leave the initial ordering unchanged: $(s,\sigma) \leq_a (s',\sigma')$ iff $s \leq_a s'$. Giving priority to action plausibility does not in any way mean that the agent's belief in actions is stronger than her belief in states; it just captures the fact that, at the time of updating with a given action, *the belief about the action is what is actual, it is the current belief about what is going on, while the beliefs about the input-states are in the past.*[19]

In a nutshell: *the doxastic action is the one that changes the initial doxastic state, and not vice-versa.* The belief update induced by a given action is nothing but an update with the (presently) believed action. If the believed action σ requires the agent to revise some past beliefs, then so be it: this is the whole point of believing σ, namely to use it to revise one's past beliefs. For example, in a successful lying, the action plausibility relation makes the hearer believe that the speaker is telling the truth; so she'll accept this message (unless contradicted by her knowledge), and change her past beliefs appropriately: this is what makes the lying successful.

Action-priority update generalizes product update. Recall the definition of the epistemic indistinguishability relation $\sim_a$ in a plausibility model: $s \sim_a s'$ iff either $s \leq_a s'$ or $s' \leq_a s$. It is easy to see that the Action Priority Update implies the familiar update rule from [6, 5], known in Dynamic Epistemic Logic as the "product update":

$$(s,\sigma) \sim_a (s',\sigma') \text{ iff } s \sim_a s' \text{ and } \sigma \sim_a \sigma'.$$

Program transitions. For every state model $\mathbf{S}$, every program $\Gamma \subseteq \Sigma$ over an action model $\mathbf{\Sigma}$ induces a transition relation $\xrightarrow{\Gamma}_{\mathbf{S}} \subseteq \mathbf{S} \times (S \otimes \Sigma)$ from $\mathbf{S}$ to $\mathbf{S} \otimes \mathbf{\Sigma}$, given by:

$$s \xrightarrow{\Gamma}_{\mathbf{S}} (s',\gamma) \text{ iff } s = s',\ (s,\gamma) \in S \otimes \Sigma \text{ and } \gamma \in \Gamma\,.$$

3.3 Simulating various belief-revision policies

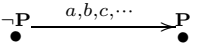
We give here three examples of *multi-agent belief-revision policies* that can be simulated by our product update: *truthful public announcements of "hard facts"*, *"lexicographic update"* and *"conservative upgrade"*. They were all introduced by van Benthem in [14], as multi-agent versions of revision operators previously considered by Rott [45] and others.

Public announcements of "hard facts". A *truthful public announcement* $!\mathbf{P}$ *of some "hard fact"* $\mathbf{P}$ is not really about belief revision, but about the learning of *certified true information*: it establishes *common knowledge*

[19] Of course, *at a later moment*, the above-mentioned belief about action (*now* belonging to the past) might be itself revised. But this is another, *future update*.

that **P** was the case. This is the action described in [14] as (public) "belief change under hard facts". As an operation on models, this is described in [14] as taking any state model **S** and *deleting all the non-P states, while keeping the same indistinguishability and plausibility relations between the surviving states*. In our setting, the corresponding action model consists of only one node, labeled with **P**. It is easy to see that the above operation on models can be exactly "simulated" by taking the anti-lexicographic product update with this one-node action model.

Public announcements of "soft facts": The "lexicographic upgrade". To allow for "soft" belief revision, an operation $\Uparrow\mathbf{P}$ was introduced in [14], essentially adapting to public announcements the 'lexicographic' policy for belief revision described in [45]. This operation, called "lexicographic update" consists of changing the current plausibility order on any given state model as follows: *every* $\mathbf{P}$*-world becomes "better" (more plausible) than all* $\neg\mathbf{P}$*-worlds in the same information cell, and within the two zones (*$\mathbf{P}$ *and* $\neg\mathbf{P}$*), the old ordering remains*. In our setting, this action corresponds to the following local plausibility action model:



Taking the anti-lexicographic update product with this action will give an exact "simulation" of the lexicographic upgrade operation.

"Conservative upgrade". The operation $\uparrow\mathbf{P}$ of "conservative upgrade", also defined in [14], changes any model as follows: *in every information cell, the best* $\mathbf{P}$*-worlds become better than all the worlds in that cell* (i.e., in every cell the most plausible **P**-states become the most plausible overall in that cell), *and apart from that, the old order remains*. In the case of a system *with only one agent*, it is easy to see that we have $\uparrow\mathbf{P} = \Uparrow(*_a\mathbf{P})$, where $*_a$ is the unary "revision modality" introduced in the previous section. In the case of a set $\mathcal{A} = \{1, \cdots, n\}$ with $n > 1$ agents, we can simulate $\uparrow\mathbf{P}$ using a model with 2^n actions $\{\uparrow_I\mathbf{P}\}_{I\subseteq\mathcal{A}}$, with

$$\mathrm{pre}_{\uparrow_I\mathbf{P}} = \bigwedge_{i\in I} *_i\mathbf{P} \wedge \bigwedge_{j\notin I} \neg *_j \mathbf{P},$$

$$\uparrow_I\mathbf{P} \leq_k \uparrow_J\mathbf{P} \quad \text{iff} \quad J \cap \{k\} \subseteq I.$$

3.4 Operations on doxastic programs

First, we introduce *dynamic modalities*, capturing *the "weakest precondition" of a program* Γ. These are the natural analogues of the PDL modalities for our program transition relations $\xrightarrow{\Gamma}$ between models.

Dynamic modalities. Let $\mathbf{\Sigma}$ be some action plausibility model and $\Gamma \subseteq \Sigma$ be a doxastic model over $\mathbf{\Sigma}$. For every doxastic proposition $\mathbf{P}$, we define a doxastic proposition $[\Gamma]\mathbf{P}$ given by

$$([\Gamma]\mathbf{P})_{\mathbf{S}} \;:=\; [\stackrel{\Gamma}{\rightarrow}_{\mathbf{S}}]\mathbf{P}_{\mathbf{S}} = \{s \in S : \forall t \in S \otimes \Sigma \, (s \stackrel{\Gamma}{\rightarrow}_{\mathbf{S}} t \;\Rightarrow\; t \models_{\mathbf{S}\otimes\mathbf{\Sigma}} \mathbf{P}) \}.$$

For *basic doxastic actions* $\sigma \in \Sigma$, we define the dynamic modality $[\sigma]$ via the above-mentioned identification of actions σ with singleton programs $\{\sigma\}$:

$$([\sigma]\mathbf{P})_{\mathbf{S}} := ([\{\sigma\}]\mathbf{P})_{\mathbf{S}} = \{s \in S : \text{ if } (s,\sigma) \in \mathbf{S} \otimes \mathbf{\Sigma} \text{ then } (s,\sigma) \in \mathbf{P}_{\mathbf{S}\otimes\mathbf{\Sigma}}\}.$$

The dual (Diamond) modalities are defined as usually: $\langle\Gamma\rangle\mathbf{P} := \neg[\Gamma]\neg\mathbf{P}$.

We can now introduce operators on doxastic programs that are the analogues of the *regular operations* of PDL.

Sequential composition. The *sequential composition* $\mathbf{\Sigma};\mathbf{\Delta}$ of two action plausibility models $\mathbf{\Sigma} = (\Sigma, \leq_a, \mathrm{pre})$, $\mathbf{\Delta} = (\Delta, \leq_a, \mathrm{pre})$ is defined as follows:

- the set of basic actions is the Cartesian product $\Sigma \times \Delta$
- the preconditions are given by $\mathrm{pre}_{(\sigma,\delta)} := \langle\sigma\rangle \, \mathrm{pre}_\delta$
- the plausibility order is given by putting $(\sigma,\delta) \leq_a (\sigma',\delta')$ iff: either $\sigma <_a \sigma'$ and $\delta \sim_a \delta'$, or else $\sigma \cong_a \sigma'$ and $\delta \leq_a \delta'$.

We think of (σ,δ) as the action of *performing first* σ *then* δ, and thus use the notation

$$\sigma;\delta := (\sigma,\delta).$$

We can extend this notation to doxastic programs, by defining the *sequential composition of programs* $\Gamma \subseteq \Sigma$ and $\Lambda \subseteq \Delta$ to be a program $\Gamma;\Lambda \subseteq \Sigma;\Delta$ over the action model $\mathbf{\Sigma};\mathbf{\Delta}$, given by:

$$\Gamma;\Lambda := \{(\gamma,\lambda) : \gamma \in \Gamma, \lambda \in \Lambda\}.$$

It is easy to see that this behaves indeed like a sequential composition:

Proposition 3.7. For every state plausibility model $\mathbf{S}$, action plausibility models $\mathbf{\Sigma}$ and $\mathbf{\Delta}$, and programs $\Gamma \subseteq \Sigma$, $\Lambda \subseteq \Delta$, we have the following:

1. The state plausibility models $(\mathbf{S} \otimes \mathbf{\Sigma}) \otimes \mathbf{\Delta}$ and $\mathbf{S} \otimes (\mathbf{\Sigma};\mathbf{\Delta})$ are isomorphic, via the canonical map $F : (S \otimes \Sigma) \otimes \Delta \to S \otimes (\Sigma;\Delta)$ given by
$$F(\,((s,\sigma),\delta)\,) \;:=\; (s,(\sigma,\delta)).$$

2. The transition relation for the program $\Gamma;\Delta$ is the relational composition of the transition relations for Γ and for Δ and of the isomorphism map F:

$$s \overset{\Gamma;\Delta}{\rightarrow}_{\mathbf{S}} s' \text{ iff there exist } w,t \in S \otimes \Sigma \text{ such that } s \overset{\Gamma}{\rightarrow}_{\mathbf{S}} w \overset{\Delta}{\rightarrow}_{\mathbf{S}\otimes\mathbf{\Sigma}} t \text{ and } F(t) = s'.$$

Union (non-deterministic choice). If $\mathbf{\Sigma} = (\Sigma, \leq_a, \mathrm{pre})$ and $\mathbf{\Delta} = (\Delta, \leq'_a, \mathrm{pre}')$ are two action plausibility models, their *disjoint union* $\mathbf{\Sigma} \sqcup \mathbf{\Delta}$ is simply given by taking as set of states the disjoint union $\Sigma \sqcup \Delta$ of the two sets of states, taking as plausibility order the disjoint union $\leq_a \sqcup \leq'_a$ and as precondition map the disjoint union $\mathrm{pre} \sqcup \mathrm{pre}'$ of the two precondition maps. If $\Gamma \subseteq \Sigma$ and $\Lambda \subseteq \Delta$ are doxastic programs over the two models, we define their *union* to be the program over the model $\mathbf{\Sigma} \sqcup \mathbf{\Delta}$ given by the disjoint union $\Gamma \sqcup \Lambda$ of the the sets of actions of the two programs.

Again, it is easy to see that *this behaves indeed like a non-deterministic choice operator*:

Proposition 3.8. Let $i_1 : \Sigma \to \Sigma \sqcup \Delta$ and $i_2 : \Delta \to \Sigma \sqcup \Delta$ be the two canonical injections. Then the following are equivalent:

- $s \overset{\Gamma\sqcup\Delta}{\rightarrow}_{\mathbf{S}} s'$
- there exists t such that:

$$\text{either } s \overset{\Gamma}{\rightarrow}_{\mathbf{S}} t \text{ and } i_1(t) = s', \text{ or else } s \overset{\Delta}{\rightarrow}_{\mathbf{S}} t \text{ and } i_2(t) = s'.$$

Other operators. *Arbitrary unions* $\bigsqcup_i \Gamma_i$ can be similarly defined, and then one can define *iteration* $\Gamma^* := \bigsqcup_i \Gamma^i$ (where $\Gamma^0 = !\top$ and $\Gamma^{i+1} = \Gamma;\Gamma^i$).

3.5 The laws of dynamic belief revision

The "laws of dynamic belief revision" are the fundamental equations of Belief Dynamics, allowing us *to compute future doxastic attitudes from past ones*, given the doxastic events that happen in the meantime. In modal terms, these can be stated as "reduction laws" for inductively computing dynamic modalities $[\Gamma]\mathbf{P}$, by reducing them to modalities $[\Gamma']\mathbf{P}'$ in which either the propositions $\mathbf{P}'$ or the programs Γ' have *lower complexity*.

The following immediate consequence of the definition of $[\Gamma]\mathbf{P}$ allows us to reduce modalities for non-deterministic programs Γ to the ones for their deterministic resolutions $\gamma \in \Gamma$:

Deterministic Resolution Law. For every program $\Gamma \subseteq \Sigma$, we have

$$[\Gamma]\mathbf{P} \;=\; \bigwedge_{\gamma\in\Gamma}[\gamma]\mathbf{P}\,.$$

So, for our other laws, we can restrict ourselves to *basic actions* in Σ.

The Action-Knowledge Law. For every action $\sigma \in \Sigma$, we have:

$$[\sigma]K_a\mathbf{P} \;=\; \mathrm{pre}_\sigma \to \bigwedge_{\sigma'\sim_a\sigma} K_a[\sigma']\mathbf{P}\,.$$

This Action-Knowledge Law is essentially the same as in [6, 5]: *a proposition* $\mathbf{P}$ *will be known after a doxastic event iff, whenever the event can take place, it is known that* $\mathbf{P}$ *will become true after all events that are indistinguishable from the given one.*

The Action-Safe-Belief Law. For every action $\sigma \in \Sigma$, we have:

$$[\sigma]\Box_a\mathbf{P} \;=\; \mathrm{pre}_\sigma \to \bigwedge_{\sigma'<_a\alpha} K_a[\sigma']\mathbf{P} \wedge \bigwedge_{\sigma''\cong_a\sigma} \Box_a[\sigma'']\mathbf{P}\,.$$

This law embodies the essence of the Action-Priority Rule: *a proposition* $\mathbf{P}$ *will be safely believed after a doxastic event iff, whenever the event can take place, it is known that* $\mathbf{P}$ *will become true after all more plausible events and in the same time it is safely believed that* $\mathbf{P}$ *will become true after all equi-plausible events.*

Since we took knowledge and safe belief as the basis of our static logic $K\Box$, the above two laws are the "fundamental equations" of our theory of dynamic belief revision. But note that, as a consequence, one can obtain *derived laws for (conditional) belief* as well. Indeed, using the above-mentioned characterization of conditional belief in terms of K and $\Box$, we obtain the following:

The Derived Law of Action-Conditional-Belief. For every action $\sigma \in \Sigma$, we have:

$$[\sigma]B_a^{\mathbf{P}}\mathbf{Q} \;=\; \mathrm{pre}_\sigma \to \bigvee_{\Gamma\subseteq\Sigma}\Big(\bigwedge_{\gamma\in\Gamma}\tilde{K}_a\langle\gamma\rangle\mathbf{P} \wedge \bigwedge_{\gamma'\notin\Gamma}\neg\tilde{K}_a\langle\gamma'\rangle\mathbf{P} \wedge B_a^{\langle\sigma_a^\Gamma\rangle\mathbf{P}}[\sigma_a^\Gamma]\mathbf{Q}\Big)\,.$$

This derived law, a version of which was first introduced in [10] (where it was considered a fundamental law), allows us to predict future conditional beliefs from current conditional beliefs.

To explain the meaning of this law, we re-state it as follows: For every $s \in \mathbf{S}$ and $\sigma \in \mathbf{\Sigma}$, we have:

$$s \models [\sigma]B_a^{\mathbf{P}}\mathbf{Q} \quad \text{iff} \quad s \models \mathrm{pre}_\sigma \to B_a^{\langle\sigma_a^\Gamma\rangle\mathbf{P}}[\sigma_a^\Gamma]\mathbf{Q}\,,$$
$$\text{where } \Gamma = \{\gamma \in \Sigma : s \models_{\mathbf{S}} \tilde{K}_a\langle\gamma\rangle\mathbf{P}\}\,.$$

It is easy to see that this "local" (state-dependent) version of the reduction law is equivalent to the previous (state-independent) one. The set Γ encodes the extra information about the current action that is given to the agent by the context s and by the post-condition $\mathbf{P}$; while σ_a^Γ is the action's *post-conditional contextual appearance*, i.e., the way it appears to the agent in the view of this extra-information Γ. Indeed, a given action might "appear" differently in a given context (i.e., at a state s) than it does in general: the information possessed by the agent at the state s might imply the negation of certain actions, hence their impossibility; this information will then be used to revise the agent's beliefs about the actions, obtaining her contextual beliefs. Moreover, in the presence of further information (a "post-condition" $\mathbf{P}$), this appearance might again be revised. The "post-conditional contextual appearance" is the result of this double revision: the agent's belief about action σ is revised with the information given to her by the context s and the post-condition $\mathbf{P}$. This information is encoded in a set $\Gamma = \{\gamma \in \Sigma : s \models_{\mathbf{S}} \tilde{K}_a\langle\gamma\rangle\mathbf{P}\}$ of "admissible" actions: the actions for which the agent considers epistemically possible (at s) that they can be performed and they can achieve the post-condition $\mathbf{P}$. The "post-conditional contextual appearance" σ_a^Γ of action σ captures the agent's revised theory about σ after revision with the relevant information Γ.

So the above law says that: *the agent's future conditional beliefs* $[\sigma]B_a^{\mathbf{P}}$ *can be predicted, given that action* σ *happens, by her current conditional beliefs* $B_a^{\langle\sigma_a^\Gamma\rangle\mathbf{P}}[\sigma_a^\Gamma]$ *about what will be true after the apparent action* σ_a^Γ (as it appears in the given context and in the view of the given post-condition $\mathbf{P}$), *beliefs conditioned on the information* $(\langle\sigma_a^\Gamma\rangle\mathbf{P})$ *that the apparent action* σ_a^Γ *actually can lead to the fulfillment of the post-condition* $\mathbf{P}$.

Special cases. As special cases of the Action-Conditional-Belief Law, we can derive *all the reduction laws* in [14] for (conditional) belief after the events $!\mathbf{P}$, $\Uparrow\mathbf{P}$ and $\uparrow\mathbf{P}$:

$$[!\,\mathbf{P}]B_a^{\mathbf{Q}}\mathbf{R} = \mathbf{P} \to B_a^{\mathbf{P}\wedge[!\,\mathbf{P}]\mathbf{Q}}[!\,\mathbf{P}]\mathbf{R}\,,$$

$$[\Uparrow\mathbf{P}]B_a^{\mathbf{Q}}\mathbf{R} = \left(\tilde{K}_a^{\mathbf{P}}[\Uparrow\mathbf{P}]\mathbf{Q} \wedge B_a^{\mathbf{P}\wedge[\Uparrow\mathbf{P}]\mathbf{Q}}[\Uparrow\mathbf{P}]\mathbf{R}\right) \vee \left(\neg\tilde{K}_a^{\mathbf{P}}[\Uparrow\mathbf{P}]\mathbf{Q} \wedge B_a^{[\Uparrow\mathbf{P}]\mathbf{Q}}[\Uparrow\mathbf{P}]\mathbf{R}\right)\,,$$

$$[\uparrow\mathbf{P}]B_a^{\mathbf{Q}}\mathbf{R} = \left(\tilde{B}_a^{\mathbf{P}}[\uparrow\mathbf{P}]\mathbf{Q} \wedge B_a^{\mathbf{P}\wedge[\uparrow\mathbf{P}]\mathbf{Q}}[\uparrow\mathbf{P}]\mathbf{R}\right) \vee \left(\neg\tilde{B}_a^{\mathbf{P}}[\uparrow\mathbf{P}]\mathbf{Q} \wedge B_a^{[\uparrow\mathbf{P}]\mathbf{Q}}[\uparrow\mathbf{P}]\mathbf{R}\right)\,,$$

where

$$K_a^{\mathbf{P}}\mathbf{Q} := K_a(\mathbf{P}\to\mathbf{Q})\,, \qquad \tilde{K}_a^{\mathbf{P}}\mathbf{Q} := \neg K_a^{\mathbf{P}}\neg\mathbf{Q}\,, \qquad \tilde{B}_a^{\mathbf{P}}\mathbf{Q} := \neg B_a^{\mathbf{P}}\neg\mathbf{Q}\,.$$

Laws for other doxastic attitudes. The *equi-plausibility modality behaves dynamically "almost" like knowledge*, while *the strict plausibility*

modality behaves like safe belief, as witnessed by the following laws:

$$[\sigma][\cong_a]\mathbf{P} \;=\; \mathrm{pre}_\sigma \to \bigwedge_{\sigma' \cong_a \sigma} [\cong_a][\sigma']\mathbf{P}\ ,$$

$$[\sigma][>_a]\mathbf{P} \;=\; \mathrm{pre}_\sigma \to \bigwedge_{\sigma' <_a \alpha} K_a[\sigma']\mathbf{P} \wedge \bigwedge_{\sigma'' \cong_a \sigma} [>_a][\sigma'']\mathbf{P}\ .$$

From these, we can *derive laws for all the other doxastic attitudes* above.

3.6 The logic of doxastic actions

The problem of finding a general syntax for action models has been tackled in various ways by different authors. Here we use the *action-signature approach* from [5].

Signature. A doxastic *action signature* is a *finite plausibility frame* $\mathbf{\Sigma}$, together with an *ordered list without repetitions* $(\sigma_1, \ldots, \sigma_n)$ of some of the elements of $\mathbf{\Sigma}$. The elements of Σ are called *action types*. A type σ is called *trivial* if it is *not* in the above list.

Example 3.9. The *"hard" public announcement signature* HardPub is a singleton frame, consisting of one action type !, identity as the order relation, and the list (!).

The *"soft" public announcement signature* SoftPub is a two-point frame, consisting of types $\Uparrow$ and $\Downarrow$, with $\Downarrow <_a \Uparrow$ for all agents a, and the list $(\Uparrow, \Downarrow)$.

Similarly, one can define the signatures of *fully private announcements with n alternatives*, *private "fair-game" announcements, conservative upgrades* etc. As we will see below, *there is no signature of "successful (public) lying": public lying actions fall under the type of "soft" public announcements*, so they are generated by that signature.

Languages. For each action signature $(\mathbf{\Sigma},\ (\sigma_1, \ldots, \sigma_n))$, the language $L(\mathbf{\Sigma})$ consists of a set of *sentences* φ and a set of *program terms* π, defined by simultaneous recursion:

$$\varphi \;::=\; p \mid \neg\varphi \mid \varphi \wedge \varphi \mid K_a\varphi \mid \Box_a\varphi \mid [\pi]\varphi$$
$$\pi \;::=\; \sigma\varphi_1 \ldots \varphi_n \mid \pi \sqcup \pi \mid \pi;\pi$$

where $p \in \Phi$, $a \in \mathcal{A}$, $\sigma \in \Sigma$, and $\sigma\varphi_1 \ldots \varphi_n$ is an expression consisting of σ and a string of n sentences, where n is the length of the list $(\sigma_1, \ldots, \sigma_n)$.

Syntactic action model. The expressions of the form $\sigma\vec{\varphi}$ are called *basic programs*. The preorders on $\mathbf{\Sigma}$ induce in a natural way preorders on the basic programs in $L(\mathbf{\Sigma})$:

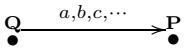
$$\sigma\vec{\varphi} \leq_a \sigma\vec{\psi} \;\text{ iff }\; \sigma \leq_a \sigma' \text{ and } \vec{\varphi} = \vec{\psi}.$$

The given listing can be used to assign syntactic preconditions for basic programs, by putting: $\text{pre}_{\sigma_i\vec{\varphi}} := \varphi_i$, and $\text{pre}_{\sigma\vec{\varphi}} := \top$ (the trivially true sentence) if σ is not in the listing. Thus, the basic programs of the form $\sigma\vec{\varphi}$ form a *"syntactic plausibility model"* $\mathbf{\Sigma}\vec{\varphi}$; i.e., every given interpretation $\|\cdot\| : L(\mathbf{\Sigma}) \to \mathsf{Prop}$ of sentences as doxastic propositions will convert this syntactic model into a "real" (semantic) plausibility model, called $\mathbf{\Sigma}\|\vec{\varphi}\|$.

Action models induced by a signature. For a given signature $\mathbf{\Sigma}$, let $(\sigma_1, \ldots, \sigma_n)$ be its list of non-trivial types, and let $\vec{\mathbf{P}} = (\mathbf{P}_1, \ldots, \mathbf{P}_n)$ be a matching list of doxastic propositions. The *action model generated by the signature* $\mathbf{\Sigma}$ *and the list of propositions* $\vec{\mathbf{P}}$ is the model $\mathbf{\Sigma}\vec{\mathbf{P}}$, having $\mathbf{\Sigma}$ as its underlying action frame and having a precondition map given by: $\text{pre}_{\sigma_i} = \mathbf{P}_i$, for non-trivial types σ_i; and $\text{pre}_{\sigma} = \top$ (the trivially true proposition), for trivial types σ. When referring to σ as an *action* in $\mathbf{\Sigma}\vec{\mathbf{P}}$, we will denote it by $\sigma\vec{\mathbf{P}}$, to distinguish it from the action *type* $\sigma \in \Sigma$.

We can obviously extend this construction to *sets of action types*: given a signature $\mathbf{\Sigma}$ and a list $\vec{\mathbf{P}} = (\mathbf{P}_1, \ldots, \mathbf{P}_n)$, every set $\Gamma \subseteq \Sigma$ gives rise to a doxastic program $\Gamma\vec{\mathbf{P}} := \{\sigma\vec{\mathbf{P}} : \sigma \in \Sigma\} \subseteq \Sigma\vec{\mathbf{P}}$.

Example 3.10. The action model of a hard public announcement $!\mathbf{P}$ is generated as $!(\mathbf{P})$ by the hard public announcement signature $\mathsf{HardPub} = \{!\}$ and the list $(\mathbf{P})$. Similarly, the action model $\mathsf{SoftPub}(\mathbf{P})$ generated by the *soft* public announcement signature $\mathsf{SoftPub}$ and a list $(\mathbf{P}, \mathbf{Q})$ of two propositions consists of two actions $\Uparrow(\mathbf{P}, \mathbf{Q})$ and $\Downarrow(\mathbf{P}, \mathbf{Q})$, with $\Uparrow(\mathbf{P}, \mathbf{Q}) <_a \Downarrow(\mathbf{P}, \mathbf{Q})$, $\text{pre}_{\Uparrow(\mathbf{P},\mathbf{Q})} = \mathbf{P}$ and $\text{pre}_{\Downarrow(\mathbf{P},\mathbf{Q})} = \mathbf{Q}$:



This represents an event during which all agents share a common belief that $\mathbf{P}$ is announced; but they might be wrong and maybe $\mathbf{Q}$ was announced instead. However, it is common knowledge that either $\mathbf{P}$ or $\mathbf{Q}$ was announced.

Successful (public) lying $\mathrm{Lie}\,\mathbf{P}$ (*by an anonymous agent, falsely announcing* $\mathbf{P}$) can now be expressed as $\mathrm{Lie}\,\mathbf{P} := \Downarrow(\mathbf{P}, \neg\mathbf{P})$. The *truthful* soft announcement is $\mathrm{True}\,\mathbf{P} := \Uparrow(\mathbf{P}, \neg\mathbf{P})$. Finally, the soft public announcement (lexicographic update) $\Uparrow\mathbf{P}$, as previously defined, is given by the nondeterministic union $\Uparrow\mathbf{P} := \mathrm{True}\,\mathbf{P} \sqcup \mathrm{Lie}\,\mathbf{P}$.

Semantics. We define by simultaneous induction two *interpretation maps*, one taking sentences φ into doxastic propositions $\|\varphi\| \in \mathsf{Prop}$, the second taking program terms π into doxastic programs $\|\pi\|$ over some plausibility frames. The inductive definition uses the obvious semantic clauses. For programs: $\|\sigma\vec{\varphi}\|$ is the action $\sigma\|\vec{\varphi}\|$ (or, more exactly, the singleton program $\{\sigma\|\vec{\varphi}\|\}$ over the frame $\mathbf{\Sigma}\|\vec{\varphi}\|$), $\|\pi \sqcup \pi'\| := \|\pi\| \sqcup \|\pi'\|$, $\|\pi; \pi'\| := \|\pi\|; \|\pi'\|$.

For sentences: $\|p\|$ is as given by the valuation, $\|\neg\varphi\| := \neg\|\varphi\|$, $\|\varphi \wedge \psi\| := \|\varphi\| \wedge \|\psi\|$, $\|K_a\varphi\| := K_a\|\varphi\|$, $\|\Box_a\varphi\| := \Box_a\|\varphi\|$, $\|[\pi]\varphi\| := [\,\|\pi\|\,]\,\|\varphi\|$.

Proof system. In addition to the axioms and rules of the logic $K\Box$, the logic $L(\mathbf{\Sigma})$ includes the following Reduction Axioms:

$$
\begin{array}{rcl}
[\alpha]p & \leftrightarrow & \mathrm{pre}_\alpha \to p \\
[\alpha]\neg\varphi & \leftrightarrow & \mathrm{pre}_\alpha \to \neg[\alpha]\varphi \\
[\alpha](\varphi \wedge \psi) & \leftrightarrow & \mathrm{pre}_\alpha \to [\alpha]\varphi \wedge [\alpha]\psi \\
[\alpha]K_a\varphi & \leftrightarrow & \mathrm{pre}_\alpha \to \bigwedge\limits_{\alpha' \sim_a \alpha} K_a[\alpha']\varphi \\
[\alpha]\Box_a\varphi & \leftrightarrow & \mathrm{pre}_\alpha \to \bigwedge\limits_{\alpha' <_a \alpha} K_a[\alpha']\varphi \wedge \bigwedge\limits_{\alpha'' \cong_a \alpha} \Box_a[\alpha'']\varphi \\
[\pi \sqcup \pi']\varphi & \leftrightarrow & [\pi]\varphi \wedge [\pi']\varphi \\
[\pi;\pi']\varphi & \leftrightarrow & [\pi]\,[\pi']\varphi
\end{array}
$$

where p is any atomic sentence, π, π' are program terms, α is a *basic* program term in $L(\mathbf{\Sigma})$, pre is the syntactic precondition map defined above, and $\sim_a, <_a, \cong_a$ are respectively the (syntactic) epistemic indistinguishability, the strict plausibility order and the equi-plausibility relation on basic programs.

Theorem 3.11. For every signature $\mathbf{\Sigma}$, the above proof system for the dynamic logic $L(\mathbf{\Sigma})$ is *complete, decidable* and *has the finite model property.* In fact, this dynamic logic has the same expressive power as the "static" logic $K\Box$ of knowledge and safe belief.

Proof (Sketch). The proof is similar to the ones in [5, 6, 25]. We use the reduction laws to inductively simplify any formula until it is reduced to a formula of the $K\Box$-logic, then use the completeness of the $K\Box$ logic. Note that this is *not* an induction on subformulas, but (as in [6]) on an appropriate notion of "complexity" ordering of formulas. Q.E.D.

4 Current and Future Work, Some Open Questions

In our papers [11, 12], we present a *probabilistic version* of the theory developed here, based on *discrete (finite) Popper-Renyi conditional probability spaces* (allowing for conditionalization on events of non-zero probability, in order to cope with non-trivial belief revisions). We consider subjective probability to be the proper notion of "degree of belief", and we investigate its relationship with the qualitative concepts developed here. We develop a probabilistic generalization of the Action Priority Rule, and show that the logics presented above are *complete for the (discrete) conditional probabilistic semantics.*

We mention here a number of open questions: (1) Axiomatize the full (static) logic of doxastic attitudes introduced in this paper. It can be easily shown that they can all be reduced to the modalities K_a, $[>_a]$ and $[\cong_a]$. There are a number of obvious axioms for the resulting logic $K[>][\cong]$ (note in particular that $[>]$ satisfies the Gödel-Löb formula!), but the completeness problem is still open. (2) Axiomatize the logic of *common safe belief and common knowledge*, and their *dynamic versions*. More generally, explore the logics obtained by adding *fixed points*, or at least "epistemic regular (PDL-like) operations" as in [15], on top of our doxastic modalities. (3) Investigate the *expressive limits* of this approach *with respect to belief-revision policies*: what policies can be simulated by our update? (4) Extend the work in [11, 12], by investigating and axiomatizing doxastic logics on *infinite* conditional probability models. (5) Extend the logics with *quantitative (probabilistic)* modal operators $B^P_{a,x}Q$ (or $\Box_{a,x}Q$) expressing that *the degree of conditional belief in Q given P* (or the *degree of safety* of the belief in Q) is at least x.

Acknowledgments

Sonja Smets' contribution to this research was made possible by the post-doctoral fellowship awarded to her by the Flemish Fund for Scientific Research. We thank Johan van Benthem for his insights and help, and for the illuminating discussions we had with him on the topic of this paper. His pioneering work on dynamic belief revision acted as the "trigger" for our own. We also thank Larry Moss, Hans van Ditmarsch, Jan van Eijck and Hans Rott for their most valuable feedback. Finally, we thank the editors and the anonymous referees for their useful suggestions and comments.

References

[1] C.E. Alchourrón, P. Gärdenfors & D. Makinson. On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

[2] G. Aucher. *A Combined System for Update Logic and Belief Revision.* Master's thesis, University of Amsterdam, 2003. *ILLC Publications* MoL-2003-03.

[3] R.J. Aumann. Interactive epistemology I: Knowledge. *International Journal of Game Theory*, 28(3):263–300, 1999.

[4] A. Baltag. A logic for suspicious players: epistemic actions and belief updates in games. *Bulletin of Economic Research*, 54(1):1–46, 2002.

[5] A. Baltag & L.S. Moss. Logics for epistemic programs. *Synthese*, 139(2):165–224, 2004.

[6] A. Baltag, L.S. Moss & S. Solecki. The logic of common knowledge, public announcements, and private suspicions. In I. Gilboa, ed., *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK 98)*, pp. 43–56. 1998.

[7] A. Baltag & M. Sadrzadeh. The algebra of multi-agent dynamic belief revision. *Electronic Notes in Theoretical Computer Science*, 157(4):37–56, 2006.

[8] A. Baltag & S. Smets. Conditional doxastic models: a qualitative approach to dynamic belief revision. *Electronic Notes in Theoretical Computer Science*, 165:5–21, 2006.

[9] A. Baltag & S. Smets. Dynamic belief revision over multi-agent plausibility models. In Bonanno et al. [21], pp. 11–24.

[10] A. Baltag & S. Smets. The logic of conditional doxastic actions: a theory of dynamic multi-agent belief revision. In S. Artemov & R. Parikh, eds., *Proceedings of ESSLLI Workshop on Rationality and Knowledge*, pp. 13–30. ESSLLI, 2006.

[11] A. Baltag & S. Smets. From conditional probability to the logic of doxastic actions. In D. Samet, ed., *Proceedings of the 11th Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pp. 52–61. UCL Presses Universitaires de Louvain, 2007.

[12] A. Baltag & S. Smets. Probabilistic dynamic belief revision. In J.F.A.K. van Benthem, S. Ju & F. Veltman, eds., *A Meeting of the Minds: Proceedings of the Workshop on Logic, Rationality and Interaction, Beijing, 2007*, vol. 8 of *Texts in Computer Science*. College Publications, London, 2007.

[13] P. Battigalli & M. Siniscalchi. Strong belief and forward induction reasoning. *Journal of Economic Theory*, 105(2):356–391, 2002.

[14] J.F.A.K. van Benthem. Dynamic logic for belief revision. *Journal of Applied Non-Classical Logics*, 17(2):129–155, 2007.

[15] J.F.A.K. van Benthem, J. van Eijck & B.P. Kooi. Logics of communication and change. *Information and Computation*, 204(11):1620–1662, 2006.

[16] J.F.A.K. van Benthem, J. Gerbrandy & B. Kooi. Dynamic update with probabilities. In Bonanno et al. [21], pp. 237–246.

[17] J.F.A.K. van Benthem & F. Liu. Dynamic logic of preference upgrade. Tech. rep., University of Amsterdam, 2004. *ILLC Publications* PP-2005-29.

[18] P. Blackburn, M. de Rijke & Y. Venema. *Modal Logic*. No. 53 in Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, Cambridge, 2001.

[19] O. Board. Dynamic interactive epistemology. *Games and Economic Behaviour*, 49(1):49–80, 2002.

[20] G. Bonanno. A simple modal logic for belief revision. *Synthese*, 147(2):193–228, 2005.

[21] G. Bonanno, W. van der Hoek & M. Wooldridge, eds. *Proceedings of the 7th Conference on Logic and the Foundations of Game and Decision Theory (LOFT7)*. University of Liverpool, 2006.

[22] H.P. van Ditmarsch. *Knowledge Games*. Ph.D. thesis, University of Groningen, 2000. *ILLC Pubications* DS-2000-06.

[23] H.P. van Ditmarsch. Descriptions of game actions. *Journal of Logic, Language and Information*, 11(3):349–365, 2002.

[24] H.P. van Ditmarsch. Prolegomena to dynamic logic for belief revision. *Synthese*, 147(2):229–275, 2005.

[25] H.P. van Ditmarsch, W. van der Hoek & B.P. Kooi. *Dynamic Epistemic Logic*, vol. 337 of *Synthese Library*. Springer, 2007.

[26] H.P. van Ditmarsch & W. Labuschagne. My beliefs about your beliefs: a case study in theory of mind and epistemic logic. *Synthese*, 155(2):191–209, 2007.

[27] N. Friedmann & J.Y. Halpern. Conditional logics of belief revision. In *Proceedings of the of 12th National Conference on Artificial Intelligence (AAAI-94). Seattle, WA, USA, July 31–August 4 1994.*, pp. 915–921. AAAI Press, Menlo Park, CA, 1994.

[28] P. Gärdenfors. *Knowledge in Flux: Modelling the Dynamics of Epistemic States*. The MIT Press.

[29] J. Gerbrandy. Dynamic epistemic logic. In L.S. Moss, J. Ginzburg & M. de Rijke, eds., *Logic, Language and Information, vol. 2*, pp. 67–84. CSLI Publications, Stanford University, 1999.

[30] J. Gerbrandy & W. Groeneveld. Reasoning about information change. *Journal of Logic, Language and Information*, 6(2):147–169, 1997.

[31] J.D. Gerbrandy. *Bisimulations on Planet Kripke*. Ph.D. thesis, University of Amsterdam, 1999. *ILLC Publications* DS-1999-01.

[32] E. Gettier. Is justified true belief knowledge? *Analysis*, 23(6):121–123, 1963.

[33] P. Gochet & P. Gribomont. Epistemic logic. In D.M. Gabbay & J. Woods, eds., *Handbook of the History of Logic*, vol. 7, pp. 99–195. Elsevier, 2006.

[34] A. Grove. Two modellings for theory change. *Journal of Philosophical Logic*, 17(2):157–170, 1988.

[35] J. Hintikka. *Knowledge and Belief*. Cornell University Press, Ithaca, NY, 1962.

[36] W. van der Hoek. Systems for knowledge and beliefs. *Journal of Logic and Computation*, 3(2):173–195, 1993.

[37] H. Katsuno & A.O. Mendelzon. On the difference between updating a knowledge base and revising it. In P. Gärdenfors, ed., *Belief Revision*, Cambridge Tracts in Theoretical Computer Science, pp. 183–203. Cambridge University Press, 1992.

[38] P. Klein. A proposed definition of propositional knowledge. *Journal of Philosophy*, 68(16):471–482, 1971.

[39] B.P. Kooi. Probabilistic dynamic epistemic logic. *Journal of Logic, Language and Information*, 12(4):381–408, 2003.

[40] K. Lehrer. *Theory of Knowledge*. Routledge, London, 1990.

[41] K. Lehrer & T. Paxson, Jr. Knowledge: Undefeated justified true belief. *Journal of Philosophy*, 66(8):225–237, 1969.

[42] J.-J.Ch. Meyer & W. van der Hoek. *Epistemic Logic for AI and Computer Science*. No. 41 in Cambridge Tracts in Theoretical Computer Science. Cambridge University Press, Cambridge, 1995.

[43] G. Pappas & M. Swain, eds. *Essays on Knowledge and Justification*. Cornell Univ. Press, Ithaca, NY, 1978.

[44] J.A. Plaza. Logics of public communications. In M.L. Emrich, M.S. Pfeifer, M. Hadzikadic & Z.W. Ras, eds., *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, pp. 201–216. 1989.

[45] H. Rott. Conditionals and theory change: revisions, expansions, and additions. *Synthese*, 81(1):91–113, 1989.

[46] H. Rott. Stability, strength and sensitivity: Converting belief into knowledge. *Erkenntnis*, 61(2–3):469–493, 2004.

[47] M. Ryan & P.-Y. Schobbens. Counterfactuals and updates as inverse modalities. *Journal of Logic, Language and Information*, 6(2):123–146, 1997.

[48] K. Segerberg. Irrevocable belief revision in Dynamic Doxastic Logic. *Notre Dame Journal of Formal Logic*, 39(3):287–306, 1998.

[49] W. Spohn. Ordinal conditional functions: a dynamic theory of epistemic states. In W.L. Harper & B. Skyrms, eds., *Causation in Decision, Belief Change, and Statistics*, vol. II, pp. 105–134. 1988.

[50] R. Stalnaker. A theory of conditionals. In N. Rescher, ed., *Studies in Logical Theory*, vol. 2 of *APQ Monograph Series*. Blackwell, 1968.

[51] R. Stalnaker. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, 12:133–163, 1996.

[52] R. Stalnaker. On logics of knowledge and belief. *Philosophical Studies*, 128(1):169–199, 2006.

[53] F.P.J.M. Voorbraak. *As Far as I Know*. Ph.D. thesis, Utrecht University, Utrecht, The Netherlands, 1993. Vol. VII in *Quaestiones Infinitae*.

[54] T. Williamson. Some philosophical aspects of reasoning about knowledge. In J. van Benthem, ed., *Proceedings of the 8th Conference on Theoretical Aspects of Rationality and Knowledge (TARK'01)*, p. 97. Morgan Kaufmann Publishers, San Francisco, 2001.

A Syntactic Approach to Rationality in Games with Ordinal Payoffs

Giacomo Bonanno

Department of Economics
University of California
Davis CA 95616-8578, United States of America
gfbonanno@ucdavis.edu

Abstract

We consider strategic-form games with ordinal payoffs and provide a syntactic analysis of common belief/knowledge of rationality, which we define axiomatically. Two axioms are considered. The first says that a player is *irrational* if she chooses a particular strategy while believing that another strategy is better. We show that common belief of this weak notion of rationality characterizes the iterated deletion of pure strategies that are strictly dominated by *pure* strategies. The second axiom says that a player is *irrational* if she chooses a particular strategy while believing that a different strategy is at least as good and she considers it possible that this alternative strategy is actually better than the chosen one. We show that common *knowledge* of this stronger notion of rationality characterizes the restriction to pure strategies of the iterated deletion procedure introduced by Stalnaker (1994). Frame characterization results are also provided.

1 Introduction

The notion of rationalizability in games was introduced independently by Bernheim [2] and Pearce [16]. A strategy of player i is said to be rational if it maximizes player i's expected payoff, given her probabilistic beliefs about the strategies used by her opponents; that is, if it can be justified by some beliefs about her opponents' strategies. If player i, besides being rational, also attributes rationality to her opponents, then she must only consider as possible strategies of her opponents that are themselves justifiable. If, furthermore, player i believes that her opponents believe that she is rational, then she must believe that her opponents justify their own choices by only considering those strategies of player i that are justifiable, and so on. The strategies of player i that can be justified in this way are called rationalizable. Rationalizability was intended to capture the notion of common belief of rationality. Bernheim and Pearce showed that a strategy is rationalizable if and only if it survives the iterated deletion of strictly

Giacomo Bonanno, Wiebe van der Hoek, Michael Wooldridge (eds.). *Logic and the Foundations of Game and Decision Theory (LOFT 7)*. Texts in Logic and Games 3, Amsterdam

dominated strategies.[1] They captured the notion of common belief of rationality only informally, that is, without making use of an epistemic framework. The first epistemic characterization of rationalizability was provided by Tan and Werlang [18] using a universal type space, rather than Kripke structures (Kripke [13]). A characterization of common belief of rationality using probabilistic Kripke structures was first provided by Stalnaker [17], although it was implicit in Brandenburger and Dekel [8]. Stalnaker also introduced a new, stronger, notion of rationalizability—which he called strong rationalizability—and showed that it corresponds to an iterated deletion procedure which is stronger than the iterated deletion of strictly dominated strategies. Stalnaker's approach is entirely semantic and uses the same notion of Bayesian rationality as Bernheim and Pearce, namely expected payoff maximization. This notion presupposes that the players' payoffs are von Neumann-Morgenstern payoffs. In contrast, in this paper we consider the larger class of strategic-form games with *ordinal* payoffs. Furthermore, we take a syntactic approach and define rationality axiomatically. We consider two axioms.

The first axiom says that a player is *irrational* if she chooses a particular strategy while believing that another strategy of hers is better. We show that common belief of this weak notion of rationality characterizes the iterated deletion of strictly dominated pure strategies. Note that, in the Bayesian approach based on von Neumann-Morgenstern payoffs, it can be shown (see Pearce [16] and Brandenburger and Dekel [8]) that a pure strategy s_i of player i is a best reply to some (possibly correlated) beliefs about the strategies of her opponents if and only if there is no *mixed* strategy of player i that strictly dominates s_i. The iterated deletion of strictly dominated strategies in the Bayesian approach thus allows the deletion of a pure strategy that is dominated by a *mixed* strategy, even though it may not be dominated by another pure strategy. Since we take a purely ordinal approach, the iterated deletion procedure that we consider only allows the removal of strategies that are dominated by *pure* strategies.

The second axiom that we consider says that a player is *irrational* if she chooses a particular strategy while believing that a different strategy is at least as good and she considers it possible that this alternative strategy is actually better than the chosen one. We show that common *knowledge* of this stronger notion of rationality characterizes the iterated deletion procedure introduced by Stalnaker [17], restricted—once again—to pure strategies.

The paper is organized as follows. In the next section we review the KD45 multi-agent logic for belief and common belief and the S5 logic for knowledge and common knowledge. In Section 3 we review the definition

[1] This characterization of rationalizability is true for two-player games and extends to n-player games only if correlated beliefs are allowed (see Brandenburger and Dekel [8]).

of strategic-form game with ordinal payoffs and the iterated deletion procedures mentioned above. In Section 4 we define game logics and introduce two axioms of rationality. In Section 5 we characterize common belief of rationality in the weaker sense and common knowledge of rationality in the stronger sense. The characterization results proved in Section 5 (Propositions 5.4 and 5.8) are not characterizations in the sense in which this expression is used in modal logic, namely characterization of axioms in terms of classes of frames (see [3, p. 125]). Thus in Section 6 we provide a reformulation of our results in terms of frame characterization. In Section 7 we discuss related literature, while Section 8 contains a summary and concluding remarks.

2 Multi-agent logics of belief and knowledge

We consider a multi-modal logic with $n+1$ operators $B_1, B_2, \ldots, B_n, B_*$ where, for $i = 1, \ldots, n$, the intended interpretation of $B_i\varphi$ is "player i believes that φ", while $B_*\varphi$ is interpreted as "it is common belief that φ". The formal language is built in the usual way (see [3] and [10]) from a countable set A of atomic propositions, the connectives $\neg$ and $\vee$ (from which the connectives $\wedge$, $\rightarrow$ and $\leftrightarrow$ are defined as usual) and the modal operators.

We denote by $\mathbf{KD45}^*_n$ the logic defined by the following axioms and rules of inference.

Axioms:

1. All propositional tautologies.
2. Axiom **K** for every modal operator: for $\Box \in \{B_1, \ldots, B_n, B_*\}$,
$$\Box\varphi \wedge \Box(\varphi \rightarrow \psi) \rightarrow \Box\psi. \qquad (\mathbf{K})$$
3. Axioms **D**, **4** and **5** for individual beliefs: for $i = 1, \ldots, n$,
$$B_i\varphi \rightarrow \neg B_i \neg\varphi, \qquad (\mathbf{D}_i)$$
$$B_i\varphi \rightarrow B_i B_i\varphi, \qquad (\mathbf{4}_i)$$
$$\neg B_i\varphi \rightarrow B_i \neg B_i\varphi. \qquad (\mathbf{5}_i)$$
4. Axioms for common belief: for $i = 1, \ldots, n$,
$$B_*\varphi \rightarrow B_i\varphi, \qquad (\mathbf{CB1})$$
$$B_*\varphi \rightarrow B_i B_*\varphi, \qquad (\mathbf{CB2})$$
$$B_*(\varphi \rightarrow B_1\varphi \wedge \cdots \wedge B_n\varphi) \rightarrow (B_1\varphi \wedge \cdots \wedge B_n\varphi \rightarrow B_*\varphi). \qquad (\mathbf{CB3})$$

Rules of Inference:

1. Modus Ponens:
$$\text{From } \varphi \text{ and } (\varphi \to \psi) \text{ infer } \psi. \qquad (\mathbf{MP})$$

2. Necessitation for every modal operator: for $\Box \in \{B_1, \ldots, B_n, B_*\}$,
$$\text{From } \varphi \text{ infer } \Box\varphi. \qquad (\mathbf{Nec})$$

We denote by $\mathbf{S5}_n^*$ the logic obtained by adding to $\mathbf{KD45}_n^*$ the following axiom:

5. Axiom **T** for individual beliefs: for $i = 1, \ldots, n$,
$$B_i\varphi \to \varphi. \qquad (\mathbf{T}_i)$$

While $\mathbf{KD45}_n^*$ is a logic for individual and common beliefs, $\mathbf{S5}_n^*$ is the logic for (individual and common) knowledge. To stress the difference between the two, when we deal with $\mathbf{S5}_n^*$ we shall denote the modal operators by K_i and K_* rather than B_i and B_*, respectively.

Note that the common belief operator does not inherit all the properties of the individual belief operators. In particular, the negative introspection axiom for common belief, $\neg B_*\varphi \to B_*\neg B_*\varphi$, is *not* a theorem of $\mathbf{KD45}_n^*$. In order to obtain it as a theorem, one needs to strengthen the logic by adding the axiom that individuals are correct in their beliefs about what is commonly believed: $B_iB_*\varphi \to B_*\varphi$. Indeed, the logic $\mathbf{KD45}_n^*$ augmented with the axiom $B_iB_*\varphi \to B_*\varphi$ *coincides* with the logic $\mathbf{KD45}_n^*$ augmented with the axiom $\neg B_*\varphi \to B_*\neg B_*\varphi$ (see [6]).

On the semantic side we consider Kripke structures $\langle \Omega, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathcal{B}_* \rangle$ where Ω is a set of states or possible worlds and, for every $j \in \{1, \ldots, n, *\}$, $\mathcal{B}_j$ is a binary relation on Ω.[2] For every $\omega \in \Omega$ and for every $j \in \{1, \ldots, n, *\}$, let $\mathcal{B}_j(\omega) = \{\omega' \in \Omega : \omega \mathcal{B}_j \omega'\}$.

Definition 2.1. A $\mathrm{D45}_n^*$ *frame* is a Kripke structure $\langle \Omega, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathcal{B}_* \rangle$ that satisfies the following properties: for all $\omega, \omega' \in \Omega$ and $i = 1, \ldots, n$

1. Seriality: $\mathcal{B}_i(\omega) \neq \varnothing$;

2. Transitivity: if $\omega' \in \mathcal{B}_i(\omega)$ then $\mathcal{B}_i(\omega') \subseteq \mathcal{B}_i(\omega)$;

3. Euclideanness: if $\omega' \in \mathcal{B}_i(\omega)$ then $\mathcal{B}_i(\omega) \subseteq \mathcal{B}_i(\omega')$;

[2] Throughout the paper we shall use the Roman font for syntactic operators (e.g., B_i and K_i) and the Calligraphic font for the corresponding semantic relations (e.g., $\mathcal{B}_i$ and $\mathcal{K}_i$).

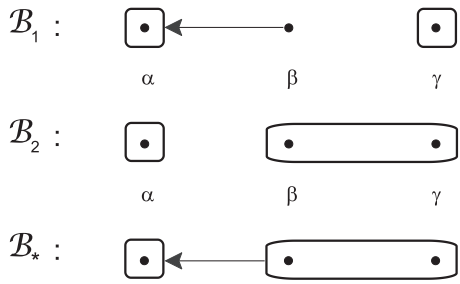


FIGURE 1. Illustration of a $D45^*_n$ frame.

4. $\mathcal{B}_*$ is the transitive closure of $\mathcal{B}_1 \cup \cdots \cup \mathcal{B}_n$, that is, $\omega' \in \mathcal{B}_*(\omega)$ if and only if there is a sequence $\langle \omega_1, \ldots, \omega_m \rangle$ in Ω such that (1) $\omega_1 = \omega$, (2) $\omega_m = \omega'$ and (3) for every $k = 1, \ldots, m-1$ there is an $i_k \in \{1, \ldots, n\}$ such that $\omega_{k+1} \in \mathcal{B}_{i_k}(\omega_k)$.

An $S5^*_n$ *frame* is a $D45^*_n$ frame that satisfies the following additional property: for all $\omega \in \Omega$ and $i = 1, \ldots, n$,

5. Reflexivity: $\omega \in \mathcal{B}_i(\omega)$.

Figure 1 illustrates the following $D45^*_n$ frame: $n = 2$, $\Omega = \{\alpha, \beta, \gamma\}$, $\mathcal{B}_1(\alpha) = \mathcal{B}_1(\beta) = \{\alpha\}$, $\mathcal{B}_1(\gamma) = \{\gamma\}$, $\mathcal{B}_2(\alpha) = \{\alpha\}$ and $\mathcal{B}_2(\beta) = \mathcal{B}_2(\gamma) = \{\beta, \gamma\}$. Thus $\mathcal{B}_*(\alpha) = \{\alpha\}$ and $\mathcal{B}_*(\beta) = \mathcal{B}_*(\gamma) = \{\alpha, \beta, \gamma\}$. We shall use the following convention when representing frames graphically: states are represented by points and for every two states ω and ω' and for every $j \in \{1, \ldots, n, *\}$, $\omega' \in \mathcal{B}_j(\omega)$ if and only if either (i) ω and ω' are enclosed in the same cell (denoted by a rounded rectangle), or (ii) there is an arrow from ω to the cell containing ω', or (iii) there is an arrow from the cell containing ω to the cell containing ω'.

The link between syntax and semantics is given by the notions of valuation and model. A $D45^*_n$ *model* (respectively, $S5^*_n$ model) is obtained by adding to a $D45^*_n$ frame (respectively, $S5^*_n$ frame) a valuation $V : A \to 2^\Omega$, where A is the set of atomic propositions and 2^Ω denotes the set of subsets of Ω. Thus a valuation assigns to every atomic proposition p the set of states where p is true. Given a model and a formula φ, we denote by $\omega \models \varphi$ the fact that φ is true at state ω. The truth set of φ is denoted by $\|\varphi\|$, that is, $\|\varphi\| = \{\omega \in \Omega : \omega \models \varphi\}$. Truth of a formula at a state is defined recursively as follows:

if $p \in A$,	$\omega \models p$ if and only if $\omega \in V(p)$,
$\omega \models \neg\varphi$	if and only if $\omega \not\models \varphi$,
$\omega \models \varphi \vee \psi$	if and only if either $\omega \models \varphi$ or $\omega \models \psi$ (or both),
$\omega \models B_i\varphi$	if and only if $\mathcal{B}_i(\omega) \subseteq \|\varphi\|$, that is,
$(i = 1, \ldots, n)$	if $\omega' \models \varphi$ for all $\omega' \in \mathcal{B}_i(\omega)$,
$\omega \models B_*\varphi$	if and only if $\mathcal{B}_*(\omega) \subseteq \|\varphi\|$.

A formula φ is valid in a model if it is true at every state, that is, if $\|\varphi\| = \Omega$. It is valid in a frame if it is valid in every model based on that frame.

The following result is well-known:[3]

Proposition 2.2. The logic $\mathbf{KD45}^*_n$ is sound and complete with respect to the class of $\mathrm{D45}^*_n$ frames, that is, a formula is a theorem of $\mathbf{KD45}^*_n$ if and only if it is valid in every $\mathrm{D45}^*_n$ frame. Similarly, $\mathbf{S5}^*_n$ is sound and complete with respect to the class of $\mathrm{S5}^*_n$ frames.

3 Ordinal games and dominance

In this paper we restrict attention to finite strategic-form (or normal-form) games with *ordinal* payoffs, which are defined as follows.

Definition 3.1. A *finite strategic-form game with ordinal payoffs* is a quintuple $G = \langle N, \{S_i\}_{i\in N}, O, \{\succeq_i\}_{i\in N}, z\rangle$, where

- $N = \{1, \ldots, n\}$ is a set of players,
- S_i is a finite set of strategies of player $i \in N$,
- O is a finite set of outcomes,
- $\succeq_i$ is player i's ordering of O,[4]
- $z : S \to O$ (where $S = S_1 \times \cdots \times S_n$) is a function that associates with every strategy profile $s = (s_1, \ldots, s_n)$ an outcome $z(s) \in O$.

Given a player i we denote by S_{-i} the set of strategy profiles of the players other than i, that is, $S_{-i} = S_1 \times \cdots \times S_{i-1} \times S_{i+1} \times \cdots \times S_n$. When we want to focus on player i we shall denote the strategy profile $s \in S$ by (s_i, s_{-i}) where $s_i \in S_i$ and $s_{-i} \in S_{-i}$.

[3] See [4]. The same result has been provided for somewhat different axiomatizations of common belief by a number of authors (for example [14], [15] and [12]).

[4] That is, $\succeq_i$ is a binary relation on O that satisfies the following properties: for all $o, o', o'' \in O$, (1) either $o \succeq_i o'$ or $o' \succeq_i o$ (completeness or connectedness) and (2) if $o \succeq_i o'$ and $o' \succeq_i o''$ then $o \succeq_i o''$ (transitivity). The interpretation of $o \succeq_i o'$ is that, according to player i, outcome o is at least as good as outcome o'. The strict ordering $\succ_i$ is defined as usual: $o \succ_i o'$ if and only if $o \succeq_i o'$ and not $o' \succeq_i o$. The interpretation of $o \succ_i o'$ is that player i strictly prefers outcome o to outcome o'.

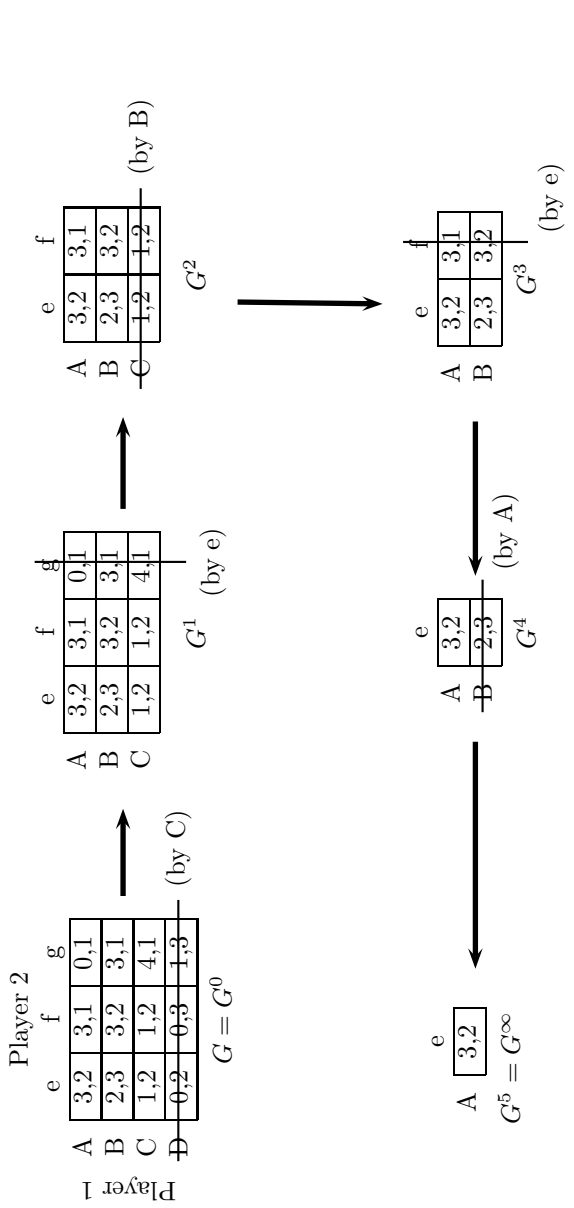


FIGURE 2. Illustration of the iterated deletion of strictly dominated strategies.

Definition 3.2. Given a game $G = \langle N, \{S_i\}_{i\in N}, O, \{\succeq_i\}_{i\in N}, z\rangle$ and $s_i \in S_i$, we say that, for player i, s_i is *strictly dominated in* G if there is another strategy $t_i \in S_i$ of player i such that—no matter what strategies the other players choose—player i prefers the outcome associated with t_i to the outcome associated with s_i, that is, if $z(t_i, s_{-i}) \succ_i z(s_i, s_{-i})$, for all $s_{-i} \in S_{-i}$.

Let $G = \langle N, \{S_i\}_{i\in N}, O, \{\succeq_i\}_{i\in N}, z\rangle$ and G' be two games, where $G' = \langle N', \{S'_i\}_{i\in N'}, O', \{\succeq'_i\}_{i\in N'}, z'\rangle$. We say that G' is a *subgame* of G if $N' = N$, $O' = O$, for every $i \in N$: $\succeq'_i = \succeq_i$ and $S'_i \subseteq S_i$ (so that $S' \subseteq S$) and z' coincides with the restriction of z to S' (that is, for every $s' \in S'$, $z'(s') = z(s')$).

Definition 3.3 (IDSDS procedure)**.** The Iterated Deletion of Strictly Dominated Strategies is the following procedure. Given a game $G = \langle N, \{S_i\}_{i\in N}, O, \{\succeq_i\}_{i\in N}, z\rangle$ let $\langle G^0, G^1, \ldots, G^m, \ldots\rangle$ be the sequence of subgames of G defined recursively as follows. For all $i \in N$,

1. let $S_i^0 = S_i$ and let $D_i^0 \subseteq S_i^0$ be the set of strategies of player i that are strictly dominated in $G^0 = G$;
2. for $m \geq 1$, let $S_i^m = S_i^{m-1} \backslash D_i^{m-1}$ and let G^m be the subgame of G with strategy sets S_i^m. Let $D_i^m \subseteq S_i^m$ be the set of strategies of player i that are strictly dominated in G^m.

Let $S_i^\infty = \bigcap_{m\in\mathbb{N}} S_i^m$ (where $\mathbb{N}$ denotes the set of non-negative integers) and let G^∞ be the subgame of G with strategy sets S_i^∞. Let $S^\infty = S_1^\infty \times \cdots \times S_n^\infty$.[5]

The IDSDS procedure is illustrated in Figure 2, where:

$$
\begin{array}{llll}
S_1^0 = \{A, B, C, D\} & D_1^0 = \{D\} & S_2^0 = \{e, f, g\} & D_2^0 = \varnothing \\
S_1^1 = \{A, B, C\} & D_1^1 = \varnothing, & S_2^1 = \{e, f, g\} & D_2^1 = \{g\} \\
S_1^2 = \{A, B, C\} & D_1^2 = \{C\} & S_2^2 = \{e, f\} & D_2^2 = \varnothing \\
S_1^3 = \{A, B\} & D_1^3 = \varnothing & S_2^3 = \{e, f\} & D_2^3 = \{f\} \\
S_1^4 = \{A, B\} & D_1^4 = \{B\} & S_2^\infty = S_2^4 = \{e\} & D_2^4 = \varnothing \\
S_1^\infty = S_1^5 = \{A\} & & &
\end{array}
$$

Thus $S^\infty = \{(A, e)\}$.

In Figure 2 we have represented the ranking $\succeq_i$ by a utility (or payoff) function $u_i : S \to \mathbb{R}$ satisfying the following property: $u_i(s) \geq u_i(s')$ if and

[5] Note that, since the strategy sets are finite, there exists an integer r such that $G^\infty = G^r = G^{r+k}$ for every $k \in \mathbb{N}$.

only if $z(s) \succeq_i z(s')$ (in each cell, the first number is the payoff of player 1 while the second number is the payoff of player 2).[6]

The next iterated deletion procedure differs from IDSDS in that at every round we delete strategy *profiles* rather than individual strategies. This procedure is the restriction to pure strategies of the algorithm introduced by Stalnaker [17].

Definition 3.4 (IDIP procedure)**.** Let $G = \langle N, \{S_i\}_{i\in N}, O, \{\succeq_i\}_{i\in N}, z\rangle$, be a game, together with a subset of strategy profiles $X \subseteq S$ and a strategy profile $x \in X$. We say that x is *inferior relative to* X if there exists a player i and a strategy $s_i \in S_i$ of player i (thus s_i need not belong to the projection of X onto S_i) such that:

1. $z(s_i, x_{-i}) \succ_i z(x_i, x_{-i})$, and
2. for all $s_{-i} \in S_{-i}$, if $(x_i, s_{-i}) \in X$ then $z(s_i, s_{-i}) \succeq_i z(x_i, s_{-i})$.

The *Iterated Deletion of Inferior Profiles* (IDIP) is defined as follows. For $m \in \mathbb{N}$ define $T^m \subseteq S$ recursively as follows: $T^0 = S$ and, for $m \geq 1$, $T^m = T^{m-1} \backslash I^{m-1}$, where $I^{m-1} \subseteq T^{m-1}$ is the set of strategy profiles that are inferior relative to T^{m-1}. Let $T^\infty = \bigcap_{m\in\mathbb{N}} T^m$.[7]

The IDIP procedure is illustrated in Figure 3, where

$$S = T^0 = \{(A,d),(A,e),(A,f),(B,d),(B,e),(B,f),(C,d),(C,e),(C,f)\},$$
$$I^0 = \{(B,e),(C,f)\}$$

(the elimination of (B,e) is done through player 2 and strategy f, while the elimination of (C,f) is done through player 1 and strategy B);

$$T^1 = \{(A,d),(A,e),(A,f),(B,d),(B,f),(C,d),(C,e)\},$$
$$I^1 = \{(B,d),(B,f),(C,e)\}$$

(the elimination of (B,d) and (B,f) is done through player 1 and strategy A, while the elimination of (C,e) is done through player 2 and strategy d);

$$T^2 = \{(A,d),(A,e),(A,f),(C,d)\},$$
$$I^2 = \{(C,d)\}$$

[6] Note that the payoff function $u_i : S \to \mathbb{R}$ used in Figure 2 to represent the ranking $\succeq_i$ of player i is an *ordinal* function in the sense that it could be replaced by any other function v_i obtained by composing u_i with a strictly increasing function on the reals. That is, if $f_i : \mathbb{R} \to \mathbb{R}$ is such that $f_i(x) > f_i(y)$ whenever $x > y$, then $v_i : S \to \mathbb{R}$ defined by $v_i(s) = f_i(u_i(s))$ could be used as an alternative representation of $\succeq_i$ and the outcome of the IDSDS algorithm would be the same.

[7] Since the strategy sets are finite, there exists an integer r such that $T^\infty = T^r = T^{r+k}$ for every $k \in \mathbb{N}$.

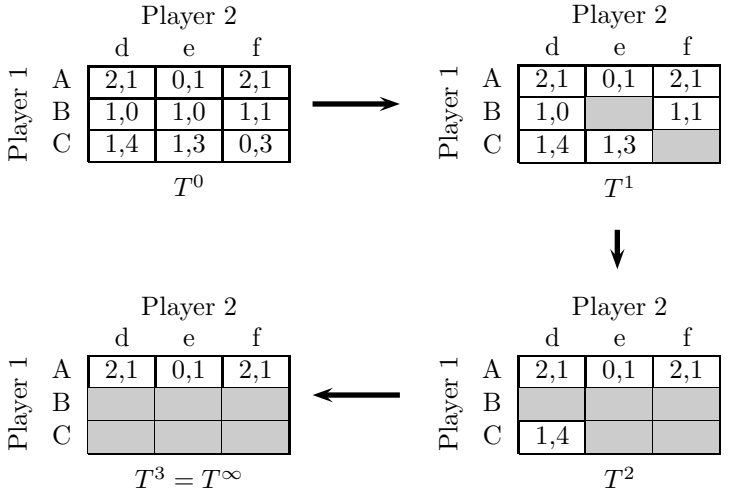


FIGURE 3. Illustration of the iterated deletion of inferior strategy profiles.

(the elimination of (C, d) is done through player 1 and strategy A);

$$T^3 = \{(A, d), (A, e), (A, f)\},$$
$$I^3 = \varnothing,$$

and thus $T^\infty = T^3$.

4 Game logics

A logic is called a *game logic* if the set of atomic propositions upon which it is built contains atomic propositions of the following form:

- Strategy symbols $s_i, t_i, \ldots$ The intended interpretation of s_i is "player i chooses strategy s_i".
- The symbols r_i whose intended interpretation is "player i is rational".
- Atomic propositions of the form $t_i \succeq_i s_i$, whose intended interpretation is "strategy t_i of player i is at least as good, for player i, as his strategy s_i", and atomic propositions of the form $t_i \succ_i s_i$, whose intended interpretation is "for player i strategy t_i is better than strategy s_i".

From now on we shall restrict attention to game logics.

Definition 4.1. Fix a game $G = \langle N, \{S_i\}_{i\in N}, O, \{\succeq_i\}_{i\in N}, z\rangle$ with $S_i = \{s_i^1, s_i^2, \ldots, s_i^{m_i}\}$ (thus the cardinality of S_i is m_i). A game logic is called a *G-logic* if its set of strategy symbols is $\{s_i^k\}_{i=1,\ldots,n;k=1,\ldots,m_i}$ (with slight abuse of notation we use the symbol s_i^k to denote both an element of S_i, that is, a strategy of player i, and an element of A, that is, an atomic proposition whose intended interpretation is "player i chooses strategy s_i^k").

Given a game G with $S_i = \{s_i^1, s_i^2, \ldots, s_i^{m_i}\}$, we denote by $\mathbf{L}_G^{\mathrm{D45}}$ (respectively, $\mathbf{L}_G^{\mathrm{S5}}$) the $\mathbf{KD45}_n^*$ (respectively, $\mathbf{S5}_n^*$) G-logic that satisfies the following additional axioms: for all $i = 1, \ldots, n$ and for all $k, \ell = 1, \ldots, m_i$, with $k \neq \ell$,

$$(s_i^1 \vee s_i^2 \vee \cdots \vee s_i^{m_i}), \tag{\textbf{G1}}$$

$$\neg(s_i^k \wedge s_i^\ell), \tag{\textbf{G2}}$$

$$s_i^k \to B_i s_i^k, \tag{\textbf{G3}}$$

$$(s_i^k \succeq_i s_i^\ell) \vee (s_i^\ell \succeq_i s_i^k), \tag{\textbf{G4}}$$

$$(s_i^\ell \succ_i s_i^k) \leftrightarrow ((s_i^\ell \succeq_i s_i^k) \wedge \neg(s_i^k \succeq_i s_i^\ell)). \tag{\textbf{G5}}$$

Axiom **G1** says that player i chooses at least one strategy, while axiom **G2** says that player i cannot choose more than one strategy. Thus **G1** and **G2** together imply that each player chooses exactly one strategy. Axiom **G3**, on the other hand, says that player i is aware of his own choice: if he chooses strategy s_i^k then he believes that he chooses s_i^k. The remaining axioms state that the ordering of strategies is complete (**G4**) and that the corresponding strict ordering is defined as usual (**G5**).

Proposition 4.2. Fix an arbitrary game G. The following is a theorem of logic $\mathbf{L}_G^{\mathrm{D45}}$: $B_i s_i^k \to s_i^k$. That is, every player has correct beliefs about her own choice of strategy.[8]

Proof. In the following PL stands for Propositional Logic. Fix a player i and $k, \ell \in \{1, \ldots, m_i\}$ with $k \neq \ell$. Let φ denote the formula

$$(s_i^1 \vee \cdots \vee s_i^{m_i}) \wedge \neg s_i^1 \wedge \cdots \wedge \neg s_i^{k-1} \wedge \neg s_i^{k+1} \wedge \cdots \wedge \neg s_i^{m_i}.$$

1.	$\varphi \to s_i^k$	tautology
2.	$\neg(s_i^k \wedge s_i^\ell)$	axiom **G2** (for $\ell \neq k$)
3.	$s_i^k \to \neg s_i^\ell$	2, PL

[8] Note that, in general, logic $\mathbf{L}_G^{\mathrm{D45}}$ allows for incorrect beliefs. In particular, a player might have incorrect beliefs about the choices made by *other* players. By Proposition 4.2, however, a player cannot have mistaken beliefs about her own choice.

4.	$B_i s_i^k \to B_i \neg s_i^\ell$	3, rule RK[9]
5.	$B_i \neg s_i^\ell \to \neg B_i s_i^\ell$	axiom $\mathbf{D}_i$
6.	$s_i^\ell \to B_i s_i^\ell$	axiom **G3**
7.	$\neg B_i s_i^\ell \to \neg s_i^\ell$	6, PL
8.	$B_i s_i^k \to \neg s_i^\ell$	4, 5, 7, PL (for $\ell \neq k$)
9.	$s_i^1 \vee \cdots \vee s_i^{m_i}$	axiom **G1**
10.	$B_i s_i^k \to (s_i^1 \vee \cdots \vee s_i^{m_i})$	9, PL
11.	$B_i s_i^k \to \varphi$	8 (for every $\ell \neq k$), 10, PL
12.	$B_i s_i^k \to s_i^k$	1, 11, PL

Q.E.D.

On the semantic side we consider models of games, which are defined as follows.

Definition 4.3. Given a game $G = \langle N, \{S_i\}_{i\in N}, O, \{\succeq_i\}_{i\in N}, z\rangle$ and a Kripke frame $F = \langle \Omega, \{\mathcal{B}_i\}_{i\in N}, \mathcal{B}_*\rangle$, a *frame for* G, or G*-frame*, is obtained by adding to F n functions $\sigma_i : \Omega \to S_i$ $(i \in N)$ satisfying the following property: if $\omega' \in \mathcal{B}_i(\omega)$ then $\sigma_i(\omega') = \sigma_i(\omega)$.

Thus a G-frame adds to a Kripke frame a function that associates with every state ω a strategy profile $\sigma(\omega) = (\sigma_1(\omega), \ldots, \sigma_n(\omega)) \in S$. The restriction that if $\omega' \in \mathcal{B}_i(\omega)$ then $\sigma_i(\omega') = \sigma_i(\omega)$ is the semantic counterpart to axiom **G3**. Given a player i, as before we will denote $\sigma(\omega)$ by $(\sigma_i(\omega), \sigma_{-i}(\omega))$, where $\sigma_{-i}(\omega) \in S_{-i}$ is the profile of strategies of the players other than i.

We say that the G-frame $\langle \Omega, \{\mathcal{B}_i\}_{i\in N}, \mathcal{B}_*, \{\sigma_i\}_{i\in N}\}\rangle$ is a $D45_n^*$ G-frame (respectively, $S5_n^*$ G-frame) if the underlying Kripke frame $\langle \Omega, \{\mathcal{B}_i\}_{i\in N}, \mathcal{B}_*\rangle$ is a $D45_n^*$ frame (respectively, $S5_n^*$ frame: see Definition 2.1).

Definition 4.4. Given a game G with $S_i = \{s_i^1, s_i^2, \ldots, s_i^{m_i}\}$, and a G-frame $F_G = \langle \Omega, \{\mathcal{B}_i\}_{i\in N}, \mathcal{B}_*, \{\sigma_i\}_{i\in N}\}\rangle$, a *model of* G, or G*-model*, is obtained by adding to F_G the following valuation:

- $\omega \models s_i^h$ if and only if $\sigma_i(\omega) = s_i^h$,
- $\omega \models (s_i^k \succeq_i s_i^\ell)$ if and only if $z(s_i^k, \sigma_{-i}(\omega)) \succeq_i z(s_i^\ell, \sigma_{-i}(\omega))$.

Thus at state ω in a G-model it is true that player i chooses strategy s_i^h if and only if the strategy of player i associated with ω is s_i^h $(\sigma_i(\omega) = s_i^h)$ and it is true that strategy s_i^k is at least as good as strategy s_i^ℓ if and only if s_i^k in combination with $\sigma_{-i}(\omega)$ (the profile of strategies of players other than i associated with ω) yields an outcome which player i considers at least as good as the outcome yielded by s_i^ℓ in combination with $\sigma_{-i}(\omega)$.

[9] RK denotes the inference rule "from $\psi \to \chi$ infer $\Box\psi \to \Box\chi$ ", which is a derived rule of inference that applies to every modal operator $\Box$ that satisfies axiom **K** and the rule of Necessitation.

	d	e	f
a	2,1	0,1	2,1
b	1,0	1,0	1,1
c	1,4	1,3	0,3

FIGURE 4. A game: player 1 controls the rows (i.e., has strategies a, b and c), and player 2 the columns.

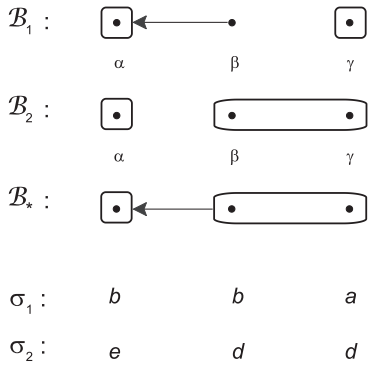


FIGURE 5. $\mathrm{D45}^*_n$ frame for the game of Figure 4.

Let $\mathbb{F}^{\mathrm{D45}}_G$ (respectively, $\mathbb{F}^{\mathrm{S5}}_G$) denote the set of $\mathrm{D45}^*_n$ (respectively, $\mathrm{S5}^*_n$) G-frames and $\mathbb{M}^{\mathrm{D45}}_G$(respectively, $\mathbb{M}^{\mathrm{S5}}_G$) the corresponding set of G-models.

Figure 4 illustrates a two-player game with strategy sets $S_1 = \{a, b, c\}$ and $S_2 = \{d, e, f\}$ and Figure 5 a $\mathrm{D45}^*_n$ frame for it. The corresponding model is given by the following valuation:

$$\begin{aligned}
\alpha &\models b \wedge e \wedge (b \succ_1 a) \wedge (c \succ_1 a) \wedge (b \succeq_1 c) \wedge (c \succeq_1 b) \\
&\qquad \wedge (f \succ_2 d) \wedge (f \succ_2 e) \wedge (e \succeq_2 d) \wedge (d \succeq_2 e), \\
\beta &\models b \wedge d \wedge (a \succ_1 b) \wedge (a \succ_1 c) \wedge (b \succeq_1 c) \wedge (c \succeq_1 b) \\
&\qquad \wedge (f \succ_2 d) \wedge (f \succ_2 e) \wedge (e \succeq_2 d) \wedge (d \succeq_2 e), \\
\gamma &\models a \wedge d \wedge (a \succ_1 b) \wedge (a \succ_1 c) \wedge (b \succeq_1 c) \wedge (c \succeq_1 b) \wedge (d \succeq_2 e) \\
&\qquad \wedge (e \succeq_2 d) \wedge (d \succeq_2 f) \wedge (f \succeq_2 d) \wedge (e \succeq_2 f) \wedge (f \succeq_2 e).
\end{aligned}$$

Proposition 4.5. Logic $\mathbf{L}^{\mathrm{D45}}_G$(respectively, $\mathbf{L}^{\mathrm{S5}}_G$) is sound with respect to the class of $\mathbb{M}^{\mathrm{D45}}_G$ (respectively, $\mathbb{M}^{\mathrm{S5}}_G$) models.

Proof. It follows from Proposition 2.2 and the following observations: (1) axioms **G1** and **G2** are valid in every model because, for every state ω there is a unique strategy $s_i^k \in S_i$ such that $\sigma_i(\omega) = s_i^k$ and, by the validation rules (see Definition 4.4), $\omega \models s_i^k$ if and only if $\sigma_i(\omega) = s_i^k$; (2) axiom **G3** is an immediate consequence of the fact (see Definition 4.3) that if $\omega' \in \mathcal{B}_i(\omega)$ then $\sigma_i(\omega') = \sigma_i(\omega)$; (3) axioms **G4** and **G5** are valid because, for every state ω, there is a unique profile of strategies $\sigma_{-i}(\omega)$ of the players other than i and the ordering $\succeq_i$ on O restricted to $z(S_i \times \sigma_{-i}(\omega))$ induces an ordering of S_i. Q.E.D.

5 Rationality and common belief of rationality

So far we have not specified what it means for a player to be rational. The first extension of $\mathbf{L}_G^{\mathrm{D45}}$ that we consider captures a very weak notion of rationality. The following axiom—called **WR** for 'Weak Rationality'—says that a player is *irrational* if she chooses a particular strategy while believing that a different strategy is better for her (recall that r_i is an atomic proposition whose intended interpretation is "player i is rational"):

$$s_i^k \wedge B_i(s_i^\ell \succ_i s_i^k) \rightarrow \neg r_i. \qquad (\mathbf{WR})$$

Given a game G, let $\mathbf{L}_G^{\mathrm{D45}}$+**WR** (respectively, $\mathbf{L}_G^{\mathrm{S5}}$+**WR**) be the extension of $\mathbf{L}_G^{\mathrm{D45}}$ (respectively, $\mathbf{L}_G^{\mathrm{S5}}$) obtained by adding axiom **WR** to it.

The next axiom—called **SR** for 'Strong Rationality'—expresses a slightly stronger notion of rationality: it says that a player is irrational if she chooses a strategy while believing that a different strategy is at least as good and she considers it possible that this alternative strategy is actually better than the chosen one:

$$s_i^k \wedge B_i(s_i^\ell \succeq_i s_i^k) \wedge \neg B_i \neg(s_i^\ell \succ_i s_i^k) \;\rightarrow\; \neg r_i. \qquad (\mathbf{SR})$$

Given a game G, let $\mathbf{L}_G^{\mathrm{D45}}$+**SR** (respectively, $\mathbf{L}_G^{\mathrm{S5}}$+**SR**) be the extension of $\mathbf{L}_G^{\mathrm{D45}}$(respectively, $\mathbf{L}_G^{\mathrm{S5}}$) obtained by adding axiom **SR** to it.

The following shows that $\mathbf{L}_G^{\mathrm{D45}}$+**SR** is an extension of $\mathbf{L}_G^{\mathrm{D45}}$+**WR**.

Proposition 5.1. **WR** is a theorem of $\mathbf{L}_G^{\mathrm{D45}}$+**SR**.

Proof. As before, PL stands for Propositional Logic.

1.	$s_i^k \wedge B_i(s_i^\ell \succeq_i s_i^k) \wedge \neg B_i \neg(s_i^\ell \succ_i s_i^k) \;\rightarrow\; \neg r_i$	Axiom **SR**
2.	$(r_i \wedge s_i^k) \rightarrow \neg(B_i(s_i^\ell \succeq_i s_i^k) \wedge \neg B_i \neg(s_i^\ell \succ_i s_i^k))$	1, PL
3.	$(s_i^\ell \succ_i s_i^k) \leftrightarrow (s_i^\ell \succeq_i s_i^k) \wedge \neg(s_i^k \succeq_i s_i^\ell)$	Axiom **G5**
4.	$(s_i^\ell \succ_i s_i^k) \rightarrow (s_i^\ell \succeq_i s_i^k)$	3, PL
5.	$B_i(s_i^\ell \succ_i s_i^k) \rightarrow B_i(s_i^\ell \succeq_i s_i^k)$	4, RK
6.	$B_i(s_i^\ell \succ_i s_i^k) \rightarrow \neg B_i \neg(s_i^\ell \succ_i s_i^k)$	Axiom $\mathbf{D}_i$
7.	$B_i(s_i^\ell \succ_i s_i^k) \rightarrow (B_i(s_i^\ell \succeq_i s_i^k) \wedge \neg B_i \neg(s_i^\ell \succ_i s_i^k))$	5, 6, PL

8. $\neg(B_i(s_i^\ell \succeq_i s_i^k) \wedge \neg B_i \neg(s_i^\ell \succ_i s_i^k)) \rightarrow \neg B_i(s_i^\ell \succ_i s_i^k)$ 7, PL
9. $(r_i \wedge s_i^k) \rightarrow \neg B_i(s_i^\ell \succ_i s_i^k)$ 2, 8, PL
10. $s_i^k \wedge B_i(s_i^\ell \succ_i s_i^k) \rightarrow \neg r_i$ 9, PL

Q.E.D.

Definition 5.2. Given a game G, let $\mathbb{M}_G^{\mathrm{D45}|\mathrm{WR}} \subseteq \mathbb{M}_G^{\mathrm{D45}}$ ($\mathbb{M}_G^{\mathrm{S5}|\mathrm{WR}} \subseteq \mathbb{M}_G^{\mathrm{S5}}$) be the class of $\mathrm{D45}_n^*$ (respectively, $\mathrm{S5}_n^*$) G-models (see Definition 4.4) where the valuation function satisfies the following additional condition:

- $\omega \models r_i$ if and only if, for every $s_i \in S_i$ there exists an $\omega' \in \mathcal{B}_i(\omega)$ such that $z(\sigma_i(\omega), \sigma_{-i}(\omega')) \succeq_i z(s_i, \sigma_{-i}(\omega'))$.[10]

Thus at state ω player i is rational if and only if, for every strategy s_i of hers, there is a state ω' that she considers possible at ω ($\omega' \in \mathcal{B}_i(\omega)$) where the strategy that she actually uses at ω ($\sigma_i(\omega)$) is at least as good as s_i against the strategies used by the other players at ω' ($\sigma_{-i}(\omega')$). For instance, in the model based on the frame of Figure 5 we have that $\alpha \models (r_1 \wedge \neg r_2)$, $\beta \models (r_1 \wedge r_2)$ and $\gamma \models (r_1 \wedge r_2)$. To see, for example, that $\beta \models r_2$ note that $\sigma_2(\beta) = d$ and for strategy f we have that $\gamma \in \mathcal{B}_2(\beta)$, $\sigma_1(\gamma) = a$ and $z(a,d) \succeq_2 z(a,f)$, while for strategy e we have that $\beta \in \mathcal{B}_2(\beta)$, $\sigma_1(\beta) = b$ and $z(b,d) \succeq_2 z(b,e)$. Thus, in the model based on the frame of Figure 5, we have that at state β both players are rational, player 2 believes that player 1 is rational, but player 1 mistakenly believes that player 2 is irrational: $\beta \models r_1 \wedge r_2 \wedge B_2 r_1 \wedge B_1 \neg r_2$.

Proposition 5.3. Logic $\mathbf{L}_G^{\mathrm{D45}}$+**WR** (respectively, $\mathbf{L}_G^{\mathrm{S5}}$+**WR**) is sound with respect to the class of models $\mathbb{M}_G^{\mathrm{D45}|\mathrm{WR}}$ (respectively, $\mathbb{M}_G^{\mathrm{S5}|\mathrm{WR}}$).

Proof. By Proposition 4.5 it is sufficient to show that axiom **WR** is valid in an arbitrary such model. Suppose that $\omega \models s_i^k \wedge B_i(s_i^\ell \succ_i s_i^k)$. Then $\sigma_i(\omega) = s_i^k$ and $\mathcal{B}_i(\omega) \subseteq \|s_i^\ell \succ_i s_i^k\|$, that is (see Definition 4.4), $z(s_i^\ell, \sigma_{-i}(\omega')) \succ_i z(s_i^k, \sigma_{-i}(\omega'))$, for every $\omega' \in \mathcal{B}_i(\omega)$. It follows from Definition 5.2 that $\omega \models \neg r_i$. Q.E.D.

The following proposition says that common belief of the weak notion of rationality expressed by axiom **WR** characterizes the Iterated Deletion of Strictly Dominated Strategies (see Definition 3.3).[11]

[10] This could alternatively be written as $z(\sigma_i(\omega'), \sigma_{-i}(\omega')) \succeq_i z(s_i, \sigma_{-i}(\omega'))$, since, by definition of G-frame (see Definition 4.3), if $\omega' \in \mathcal{B}_i(\omega)$ then $\sigma_i(\omega') = \sigma_i(\omega)$.

[11] Proposition 5.4 is the syntactic-based, ordinal version of a semantic, probabilistic-based result of Stalnaker [17]. As noted in the Introduction, Stalnaker's result was, in turn, a reformulation of earlier results due to Bernheim [2], Pearce [16], Tan and Werlang [18] and Brandenburger and Dekel [8].

The characterization results given in Propositions 5.4 and 5.8 are not characteriza-

Proposition 5.4. Fix a finite strategic-form game with ordinal payoffs G. Then both (A) and (B) below hold.

(A) Fix an arbitrary model in $\mathbb{M}_G^{\text{D45}|\text{WR}}$ and an arbitrary state ω. If $\omega \models B_*(r_1 \wedge \cdots \wedge r_n)$ then $\sigma(\omega) \in S^\infty$.

(B) For every $s \in S^\infty$ there exists a model in $\mathbb{M}_G^{\text{S5}|\text{WR}}$ and a state ω such that (1) $\sigma(\omega) = s$ and (2) $\omega \models K_*(r_1 \wedge \cdots \wedge r_n)$.[12]

Proof. (A) Fix a model in $\mathbb{M}_G^{\text{D45}|\text{WR}}$ and a state α and suppose that $\alpha \models B_*(r_1 \wedge \cdots \wedge r_n)$. The proof is by induction. First we show that, for every player $i = 1, \ldots, n$ and for every $\omega \in \mathcal{B}_*(\alpha)$, $\sigma_i(\omega) \notin D_i^0$ (see Definition 3.3). Suppose not. Then there exist a player i and a $\beta \in \mathcal{B}_*(\alpha)$ such that $\sigma_i(\beta) \in D_i^0$, that is, strategy $\sigma_i(\beta)$ of player i is strictly dominated in G by some other strategy $\hat{s}_i \in S_i$: for every $s_{-i} \in S_{-i}$, $z(\hat{s}_i, s_{-i}) \succ_i z(\sigma_i(\beta), s_{-i})$. Then, for every $\omega \in \mathcal{B}_i(\beta)$, $z(\hat{s}_i, \sigma_{-i}(\omega)) \succ_i z(\sigma_i(\beta), \sigma_{-i}(\omega))$. It follows from Definition 5.2 that $\beta \models \neg r_i$, contradicting the hypothesis that $\beta \in \mathcal{B}_*(\alpha)$ and $\alpha \models B_* r_i$. Since, for every $\omega \in \Omega$, $\sigma_i(\omega) \in S_i^0 = S_i$, it follows that, for every $\omega \in \mathcal{B}_*(\alpha)$, $\sigma_i(\omega) \in S_i^0 \backslash D_i^0 = S_i^1$. Next we prove the inductive step. Fix an integer $m \geq 1$ and suppose that, for every player $j = 1, \ldots, n$ and for every $\omega \in \mathcal{B}_*(\alpha)$, $\sigma_j(\omega) \in S_j^m$. We want to show that, for every player $i = 1, \ldots, n$ and for every $\omega \in \mathcal{B}_*(\alpha)$, $\sigma_i(\omega) \notin D_i^m$. Suppose not. Then there exist a player i and a $\beta \in \mathcal{B}_*(\alpha)$ such that $\sigma_i(\beta) \in D_i^m$, that is, strategy $\sigma_i(\beta)$ is strictly dominated in G^m by some other strategy $\tilde{s}_i \in S_i^m$. Since, by hypothesis, for every player j and for every $\omega \in \mathcal{B}_*(\alpha)$, $\sigma_j(\omega) \in S_j^m$, it follows—since $\mathcal{B}_i(\beta) \subseteq \mathcal{B}_*(\beta) \subseteq \mathcal{B}_*(\alpha)$ (see Definition 2.1)—that for every $\omega \in \mathcal{B}_i(\beta)$, $z(\tilde{s}_i, \sigma_{-i}(\omega)) \succ_i z(\sigma_i(\beta), \sigma_{-i}(\omega))$. Thus, by Definition 5.2, $\beta \models \neg r_i$, contradicting the fact that $\beta \in \mathcal{B}_*(\alpha)$ and $\alpha \models B_* r_i$. Thus, for every player $i = 1, \ldots, n$ and for every $\omega \in \mathcal{B}_*(\alpha)$, $\sigma_i(\omega) \in \bigcap_{m \in \mathbb{N}} S_i^m = S_i^\infty$. It only remains to show that $\sigma_i(\alpha) \in S_i^\infty$. Fix an arbitrary $\beta \in \mathcal{B}_i(\alpha)$. Since $\mathcal{B}_i(\alpha) \subseteq \mathcal{B}_*(\alpha)$, $\beta \in \mathcal{B}_*(\alpha)$. Thus $\sigma_i(\beta) \in S_i^\infty$. By Definition 4.3, $\sigma_i(\beta) = \sigma_i(\alpha)$. Thus $\sigma_i(\alpha) \in S_i^\infty$.

(B) Let m be the cardinality of $S^\infty = S_1^\infty \times \cdots \times S_n^\infty$ and let $\Omega = \{\omega_1, \ldots, \omega_m\}$. Let $\sigma : \Omega \to S^\infty$ be a one-to-one function. For every player i,

tions in the sense in which this expression is used in modal logic, namely characterization of axioms in terms of classes of frames (see [3, p. 125]). In Section 6 we provide a reformulation of Propositions 5.4 and 5.8 in terms of frame characterization.

[12] Recall that, in order to emphasize the distinction between belief and knowledge, when dealing with the latter we denote the modal operators by K_i and K_* rather than B_i and B_*, respectively. Similarly, we shall denote the accessibility relations by $\mathcal{K}_i$ and $\mathcal{K}_*$ rather than $\mathcal{B}_i$ and $\mathcal{B}_*$, respectively.

Thus while part (A) says that if at a state there is common *belief* of rationality then the strategy profile played at that state belongs to the set of strategy profiles that are obtained by applying the IDSDS algorithm, part (B) says that any such strategy profile is realized at a state of some model where there is common *knowledge* of rationality (that is, common belief with the added property that individual beliefs satisfy the Truth Axiom $\mathbf{T}_i$).

define the following equivalence relation on Ω: $\omega \mathcal{K}_i \omega'$ if and only if $\sigma_i(\omega) = \sigma_i(\omega')$, where $\sigma_i(\omega)$ is the ith coordinate of $\sigma(\omega)$. Let $\mathcal{K}_*$ be the transitive closure of $\bigcup_{i \in N} \mathcal{K}_i$ (then, for every $\omega \in \Omega$, $\mathcal{K}_*(\omega) = \Omega$). The structure so defined is clearly an $\mathrm{S5}^*_n$ G-frame. Consider the model corresponding to this frame (see Definition 4.4). Fix an arbitrary state ω and an arbitrary player i. By definition of S^∞, for every $s_i \in S_i$ there exists an $\omega' \in \mathcal{K}_i(\omega)$ such that $z(\sigma_i(\omega), \sigma_{-i}(\omega')) \succeq_i z(s_i, \sigma_{-i}(\omega'))$. Thus $\omega \models r_i$ (see Definition 5.2). Hence, for every $\omega \in \Omega$, $\omega \models (r_1 \wedge \cdots \wedge r_n)$ and, therefore, for every $\omega \in \Omega$, $\omega \models K_*(r_1 \wedge \cdots \wedge r_n)$. Q.E.D.

Remark 5.5. Since $\mathbb{M}_G^{\mathrm{S5|WR}} \subseteq \mathbb{M}_G^{\mathrm{D45|WR}}$ it follows from part (B) of Proposition 5.4 that the implications of common *belief* of rationality—as implicitly defined by axiom **WR**—are the same as the implications of common *knowledge* of rationality.

The above observation is not true for the stronger notion of rationality expressed by axiom **SR**, to which we now turn.

Definition 5.6. Given a game G, let $\mathbb{M}_G^{\mathrm{D45|SR}} \subseteq \mathbb{M}_G^{\mathrm{D45}}$ ($\mathbb{M}_G^{\mathrm{S5|SR}} \subseteq \mathbb{M}_G^{\mathrm{S5}}$, respectively) be the class of D45 (respectively, S5) G-models where the valuation function satisfies the following condition:

- $\omega \models r_i$ if and only if, for every $s_i \in S_i$, whenever there exists an $\omega' \in \mathcal{B}_i(\omega)$ such that $z(s_i, \sigma_{-i}(\omega')) \succ_i z(\sigma_i(\omega), \sigma_{-i}(\omega'))$ then there exists an $\omega'' \in \mathcal{B}_i(\omega)$ such that $z(\sigma_i(\omega), \sigma_{-i}(\omega'')) \succ_i z(s_i, \sigma_{-i}(\omega''))$.

Thus, at state ω, player i is rational if, whenever there is a strategy s_i of hers which is better than $\sigma_i(\omega)$ (the strategy she is actually using at ω) at some state ω' that she considers possible at ω, then $\sigma_i(\omega)$ is better than s_i at some other state ω'' that she considers possible at ω. For example, in the model based on the frame of Figure 5 we have that $\omega \models (r_1 \wedge \neg r_2)$ for every $\omega \in \{\alpha, \beta, \gamma\}$. At state β, for instance, player 2 is choosing strategy d when there is another strategy of hers, namely f, which is better than d at β and as good as d at γ, and $\mathcal{B}_2(\beta) = \{\beta, \gamma\}$. Thus she is irrational according to Definition 5.6.

It is easily verified that $\mathbb{M}_G^{\mathrm{D45|SR}} \subseteq \mathbb{M}_G^{\mathrm{D45|WR}}$ and, similarly, it is the case that $\mathbb{M}_G^{\mathrm{S5|SR}} \subseteq \mathbb{M}_G^{\mathrm{S5|WR}}$.

Proposition 5.7. Logic $\mathbf{L}_G^{\mathrm{D45}}$+**SR** (respectively, $\mathbf{L}_G^{\mathrm{S5}}$+**SR**) is sound with respect to the class of models $\mathbb{M}_G^{\mathrm{D45|SR}}$ (respectively, $\mathbb{M}_G^{\mathrm{S5|SR}}$).

Proof. By Proposition 4.5 it is sufficient to show that axiom **SR** is valid in an arbitrary such model. Suppose that $\omega \models s_i^k \wedge B_i(s_i^\ell \succeq_i s_i^k) \wedge \neg B_i \neg (s_i^\ell \succ_i s_i^k)$. Then $\sigma_i(\omega) = s_i^k$ and $\mathcal{B}_i(\omega) \subseteq \|s_i^\ell \succeq_i s_i^k\|$ [that is—see Definition 4.4—$z(s_i^\ell, \sigma_{-i}(\omega')) \succeq_i z(s_i^k, \sigma_{-i}(\omega'))$, for every $\omega' \in \mathcal{B}_i(\omega)$] and there is an $\omega'' \in$

$\mathcal{B}_i(\omega)$ such that $\omega'' \models s_i^\ell \succ_i s_i^k$, that is, $z(s_i^\ell, \sigma_{-i}(\omega'')) \succ_i z(s_i^k, \sigma_{-i}(\omega''))$. It follows from Definition 5.6 that $\omega \models \neg r_i$. Q.E.D.

The following proposition says that common *knowledge* of the stronger notion of rationality expressed by axiom **SR** characterizes the Iterated Deletion of Inferior Profiles (see Definition 3.4).[13]

Proposition 5.8. Fix a finite strategic-form game with ordinal payoffs G. Then both (A) and (B) below hold.

(A) Fix an arbitrary model in $\mathbb{M}_G^{\text{S5}|\text{SR}}$ and an arbitrary state ω. If $\omega \models K_*(r_1 \wedge \cdots \wedge r_n)$ then $\sigma(\omega) \in T^\infty$.

(B) For every $s \in T^\infty$ there exists a model in $\mathbb{M}_G^{\text{S5}|\text{SR}}$ and a state ω such that (1) $\sigma(\omega) = s$ and (2) $\omega \models K_*(r_1 \wedge \cdots \wedge r_n)$.

Proof. (A) As in the case of Proposition 5.4, the proof is by induction. Fix a model in $\mathbb{M}_G^{\text{S5}|\text{SR}}$ and a state α and suppose that $\alpha \models K_*(r_1 \wedge \cdots \wedge r_n)$. First we show that, for every $\omega \in \mathcal{K}_*(\alpha)$, $\sigma(\omega) \notin I^0$ (see Definition 3.4). Suppose, by contradiction, that there exists a $\beta \in \mathcal{K}_*(\alpha)$ such that $\sigma(\beta) \in I^0$, that is, $\sigma(\beta)$ is inferior relative to the entire set of strategy profiles S. Then there exists a player i and a strategy $\hat{s}_i \in S_i$ such that $z(\hat{s}_i, \sigma_{-i}(\beta)) \succ_i z(\sigma_i(\beta), \sigma_{-i}(\beta))$, and, for every $s_{-i} \in S_{-i}$, $z(\hat{s}_i, s_{-i}) \succeq_i z(\sigma_i(\beta), s_{-i})$. Thus $z(\hat{s}_i, \sigma_{-i}(\omega)) \succeq_i z(\sigma_i(\beta), \sigma_{-i}(\omega))$, for every $\omega \in \mathcal{K}_i(\beta)$; furthermore, by reflexivity of $\mathcal{K}_i$ (see Definition 2.1), $\beta \in \mathcal{K}_i(\beta)$. It follows from Definition 5.6 that $\beta \models \neg r_i$. Since $\beta \in \mathcal{K}_*(\alpha)$, this contradicts the hypothesis that $\alpha \models K_* r_i$. Thus, since, for every $\omega \in \Omega$, $\sigma(\omega) \in S = T^0$ we have shown that, for every $\omega \in \mathcal{K}_*(\alpha)$, $\sigma(\omega) \in T^0 \backslash I^0 = T^1$.

Now we prove the inductive step. Fix an integer $m \geq 1$ and suppose that, for every $\omega \in \mathcal{K}_*(\alpha)$, $\sigma(\omega) \in T^m$. We want to show that, for every $\omega \in \mathcal{K}_*(\alpha)$, $\sigma(\omega) \notin I^m$. Suppose, by contradiction, that there exists a $\beta \in \mathcal{K}_*(\alpha)$ such that $\sigma(\beta) \in I^m$, that is, $\sigma(\beta)$ is inferior relative to T^m. Then there exists a player i and a strategy $\tilde{s}_i \in S_i$ such that $z(\tilde{s}_i, \sigma_{-i}(\beta)) \succ_i z(\sigma_i(\beta), \sigma_{-i}(\beta))$, and, for every $s_{-i} \in S_{-i}$, if $(\tilde{s}_i, s_{-i}) \in T^m$ then $z(\tilde{s}_i, s_{-i}) \succeq_i z(\sigma_i(\beta), s_{-i})$. By Definition 4.3, for every $\omega \in \mathcal{K}_i(\beta)$, $\sigma_i(\omega) = \sigma_i(\beta)$ and by the induction hypothesis, for every $\omega \in \mathcal{K}_*(\alpha)$, $(\sigma_i(\omega), \sigma_{-i}(\omega)) \in T^m$. Thus, since $\mathcal{K}_i(\beta) \subseteq \mathcal{K}_*(\beta) \subseteq \mathcal{K}_*(\alpha)$, we have that, for every $\omega \in \mathcal{K}_i(\beta)$, $(\sigma_i(\beta), \sigma_{-i}(\omega)) \in T^m$. By reflexivity of $\mathcal{K}_i$, $\beta \in \mathcal{K}_i(\beta)$. It follows from Definition 5.6 that $\beta \models \neg r_i$. Since $\beta \in \mathcal{K}_*(\alpha)$, this contradicts the hypothesis that $\alpha \models K_* r_i$.

Thus, we have shown by induction that, for every $\omega \in \mathcal{K}_*(\alpha)$, $\sigma(\omega) \in \bigcap_{m \in \mathbb{N}} T^m = T^\infty$. It only remains to establish that $\sigma(\alpha) \in T^\infty$, but this follows from reflexivity of $\mathcal{K}_*$.

[13] Proposition 5.8 is the syntactic-based, ordinal version of a semantic, probabilistic-based result due to Stalnaker [17]. For a correction of that result see Bonanno and Nehring [5].

	c	d
a	1,1	1,0
b	1,1	0,1

FIGURE 6. A game where player 1 has strategies a and b, and player 2 has c and d.

(B) Let m be the cardinality of T^∞ and let $\Omega = \{\omega_1, \ldots, \omega_m\}$. Let $\sigma : \Omega \to T^\infty$ be a one-to-one function. For every player i, define the following equivalence relation on Ω: $\omega \mathcal{K}_i \omega'$ if and only if $\sigma_i(\omega) = \sigma_i(\omega')$, where $\sigma_i(\omega)$ is the ith coordinate of $\sigma(\omega)$. Let $\mathcal{K}_*$ be the transitive closure of $\bigcup_{i \in N} \mathcal{K}_i$ (then, for every $\omega \in \Omega$, $\mathcal{K}_*(\omega) = \Omega$). The structure so defined is clearly an $S5^*_n$ G-frame. Consider the model corresponding to this frame (see Definition 4.4). Fix an arbitrary state ω and an arbitrary player i. By definition of T^∞, for every player i and every $s_i \in S_i$ if there exists an $\omega' \in \mathcal{K}_i(\omega)$ such that if $z(s_i, \sigma_{-i}(\omega')) \succ_i z(\sigma_i(\omega), \sigma_{-i}(\omega'))$ then there exists an $\omega'' \in \mathcal{K}_i(\omega)$ such that $z(\sigma_i(\omega), \sigma_{-i}(\omega'')) \succ_i z(s_i, \sigma_{-i}(\omega''))$. Thus $\omega \models r_i$ (see Definition 5.6). Hence, for every $\omega \in \Omega$, $\omega \models (r_1 \wedge \cdots \wedge r_n)$ and, therefore, for every $\omega \in \Omega$, $\omega \models K_*(r_1 \wedge \cdots \wedge r_n)$. Q.E.D.

Note that Proposition 5.8 is not true if one replaces knowledge with belief, as illustrated in the game of Figure 6 and corresponding frame in Figure 7. In the corresponding model we have that, according to the stronger notion of rationality expressed by Definition 5.6, $\alpha \models r_1 \wedge r_2$ and $\beta \models r_1 \wedge r_2$, so that $\alpha \models B_*(r_1 \wedge r_2)$, despite the fact that $\sigma(\alpha) = (b, d)$, which is an inferior strategy profile (relative to the entire game).[14] In other words, common belief of rationality, as expressed by axiom **SR**, is compatible with the players collectively choosing an inferior strategy profile. Thus, unlike the weaker notion expressed by axiom **WR** (see Remark 5.5), with axiom **SR** there is a crucial difference between the implications of common *belief* of rationality and those of common *knowledge* of rationality.

6 Frame characterization

The characterization results proved in the previous section (Propositions 5.4 and 5.8) are not characterizations in the sense in which this expression is used in modal logic, namely characterization of axioms in terms of classes of frames (see [3, p. 125]). In this section we provide a reformulation of our results in terms of frame characterizations.

Definition 6.1. An axiom characterizes (or is characterized by) a class $\mathbb{F}$

[14] In the game of Figure 6 we have that $S^\infty = S = \{(a, c), (a, d), (b, c), (b, d)\}$ while $T^\infty = \{(a, c), (b, c)\}$.

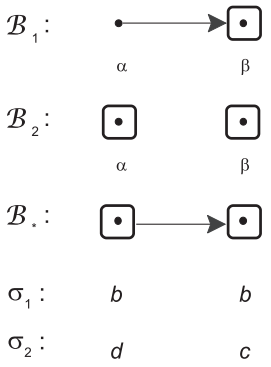


FIGURE 7. A frame for the game of Figure 6.

of Kripke frames if the axiom is valid in every model based on a frame that belongs to $\mathbb{F}$ and, conversely, if a frame does not belong to $\mathbb{F}$ then there is a model based on that frame and a state in that model at which an instance of the axiom is falsified.[15]

We now modify the previous analysis as follows. First of all, we drop the symbols r_i from the set of atomic propositions and correspondingly drop the definitions of the classes of models $\mathbb{M}_G^{\text{D45|WR}}$, $\mathbb{M}_G^{\text{S5|WR}}$, $\mathbb{M}_G^{\text{D45|SR}}$ and $\mathbb{M}_G^{\text{S5|SR}}$ (Definitions 5.2 and 5.6). Secondly we modify axioms **WR** and **SR** as follows:

$$s_i^k \to \neg B_i(s_i^\ell \succ_i s_i^k), \qquad (\mathbf{WR}')$$

$$s_i^k \to \neg(B_i(s_i^\ell \succeq_i s_i^k) \wedge \neg B_i \neg(s_i^\ell \succ_i s_i^k)). \qquad (\mathbf{SR}')$$

One can derive axioms **WR′** and **SR′** from the logics considered previously by adding the axiom that players are rational. In fact, from r_i and **WR** one obtains **WR′** (using Modus Ponens) and similarly for **SR′**.

The next proposition is the counterpart of Proposition 5.4.

Proposition 6.2. Subject to the valuation rules specified in Definition 4.4 for the atomic propositions s_i^k and $(s_i^\ell \succeq_i s_i^k)$, axiom **WR′** is characterized by the class of D45_n^* game frames (see Definition 4.3) that satisfy the following property: for all $i \in N$ and for all $\omega \in \Omega$, $\sigma_i(\omega) \in S_i^\infty$.

Proof. Fix a model based on a frame in this class, a state α, a player i and two strategies s_i^k and s_i^ℓ of player i. Suppose that $\alpha \models s_i^k$, that is,

[15] For example, as is well known, the axiom $B_i\varphi \to B_iB_i\varphi$ is characterized by the class of frames where the relation $\mathcal{B}_i$ is transitive.

$\sigma_i(\alpha) = s_i^k$. We want to show that $\alpha \models \neg B_i(s_i^\ell \succ_i s_i^k)$. Suppose not. Then $\mathcal{B}_i(\alpha) \subseteq \|s_i^\ell \succ_i s_i^k\|$, that is,

$$\text{for every } \omega \in \mathcal{B}_i(\alpha), \quad z(s_i^\ell, \sigma_{-i}(\omega)) \succ_i z(s_i^k, \sigma_{-i}(\omega)). \tag{6.1}$$

By hypothesis, for every player $j \neq i$ and for every $\omega \in \Omega$, $\sigma_j(\omega) \in S_j^\infty$. Thus it follows from this and (6.1) that $s_i^k \notin S_i^\infty$, contradicting the hypotheses that $\sigma_i(\alpha) = s_i^k$ and $\sigma_i(\omega) \in S_i^\infty$ for all $\omega \in \Omega$.

Conversely, fix a D45_n^* frame not in the class, that is, there is a state $\omega \in \Omega$ and a player $i \in N$ such that $\sigma_i(\omega) \notin S_i^\infty$. For every state ω and every player j let

$$m(\omega, j) = \begin{cases} \infty & \text{if } \sigma_j(\omega) \in S_j^\infty, \\ m & \text{if } \sigma_j(\omega) \in D_j^m. \end{cases}$$

Let $\hat{m} = \min\{m(\omega, j) : j \in N, \omega \in \Omega\}$. By our hypothesis about the frame, $\hat{m} \in \mathbb{N}$. Let $i \in N$ and $\alpha \in \Omega$ be such that $\hat{m} = m(\alpha, i)$. Then

$$\sigma_i(\alpha) \in D_i^{\hat{m}} \tag{6.2}$$

and, since (see Definition 3.3), for every $j \in N$ and for every $p, q \in \mathbb{N} \cup \{\infty\}$, $S_j^{p+q} \subseteq S_j^p$,

$$\text{for every } j \in N \text{ and } \omega \in \Omega, \quad \sigma_j(\omega) \in S_j^{\hat{m}}. \tag{6.3}$$

Let $s_i^k = \sigma_i(\alpha)$. By (6.2) and (6.3) , there exists a $s_i^\ell \in S_i$ such that, for every $\omega \in \Omega$, $\quad z(s_i^\ell, \sigma_{-i}(\omega)) \succ_i z(s_i^k, \sigma_{-i}(\omega))$. Thus $\mathcal{B}_i(\alpha) \subseteq \|s_i^\ell \succ_i s_i^k\|$ and thus $\alpha \models s_i^k \wedge B_i(s_i^\ell \succ_i s_i^k)$, so that axiom $\mathbf{WR'}$ is falsified at α. Q.E.D.

The next proposition is the counterpart of Proposition 5.8.

Proposition 6.3. Subject to the valuation rules specified in Definition 4.4 for the atomic propositions s_i^k and $(s_i^\ell \succeq_i s_i^k)$, axiom $\mathbf{SR'}$ is characterized by the class of S5_n^* game frames (see Definition 4.3) that satisfy the following property: for all $\omega \in \Omega$, $\sigma(\omega) \in T^\infty$.

Proof. Fix a model based on a frame in this class, a state α, a player i and two strategies s_i^k and s_i^ℓ of player i. Suppose that $\alpha \models s_i^k \wedge K_i(s_i^\ell \succeq_i s_i^k)$. Then $\sigma_i(\alpha) = s_i^k$ and $\mathcal{K}_i(\alpha) \subseteq \|s_i^\ell \succeq_i s_i^k\|$, that is,

$$\text{for all } \omega \in \mathcal{K}_i(\alpha), \quad z(s_i^\ell, \sigma_{-i}(\omega)) \succeq_i z(s_i^k, \sigma_{-i}(\omega)). \tag{6.4}$$

We want to show that $\alpha \models K_i \neg(s_i^\ell \succ_i s_i^k)$. Suppose not. Then there exists a $\beta \in \mathcal{K}_i(\alpha)$ such that $\beta \models (s_i^\ell \succ_i s_i^k)$, that is,

$$z(s_i^\ell, \sigma_{-i}(\beta)) \succ_i z(s_i^k, \sigma_{-i}(\beta)). \tag{6.5}$$

It follows from (6.4) and (6.5) that $(s_i^k, \sigma_{-i}(\beta)) = (\sigma_i(\beta), \sigma_{-i}(\beta))$ is inferior relative to the set $\{s \in S : s = \sigma(\omega) \text{ for some } \omega \in \mathcal{K}_i(\alpha)\}$, contradicting the hypothesis that $\sigma(\omega) \in T^\infty$ for all $\omega \in \Omega$.

Conversely, fix an $\mathrm{S5}_n^*$ frame not in the class, that is, there is a state $\omega \in \Omega$ such that $\sigma(\omega) \notin T^\infty$. For every $\omega \in \Omega$, let

$$m(\omega) = \begin{cases} \infty & \text{if } \sigma(\omega) \in T^\infty, \\ m & \text{if } \sigma(\omega) \in I^m = T^m \setminus T^{m+1}. \end{cases}$$

Let $m_0 = \min\{m(\omega) : \omega \in \Omega\}$. By our hypothesis about the frame, $m_0 \in \mathbb{N}$. Let $\alpha \in \Omega$ be such that $m_0 = m(\alpha)$. Then $\sigma(\alpha) \in I^{m_0}$, that is, there is a player i and a strategy $s_i^\ell \in S_i$ such that

$$z(s_i^\ell, \sigma_{-i}(\alpha)) \succ_i z(\sigma_i(\alpha), \sigma_{-i}(\alpha)) \tag{6.6}$$

and

$$\forall \omega \in \Omega, \text{ if } (\sigma_i(\alpha), \sigma_{-i}(\omega)) \in T^{m_0} \text{ then } z(s_i^\ell, \sigma_{-i}(\omega)) \succeq_i z(\sigma_i(\alpha), \sigma_{-i}(\omega)). \tag{6.7}$$

By definition of m_0, since (see Definition 3.4) for every $p, q \in \mathbb{N} \cup \{\infty\}$, $T^{p+q} \subseteq T^p$, for every $\omega \in \Omega$, $\sigma(\omega) \in T^{m_0}$. Thus, letting $s_i^k = \sigma_i(\alpha)$, it follows from (6.7) that $\mathcal{K}_i(\alpha) \subseteq \|s_i^\ell \succeq_i s_i^k\|$, that is, $\alpha \models K_i(s_i^\ell \succeq_i s_i^k)$. Since the frame is an S5 frame, $\mathcal{K}_i$ is reflexive and, therefore, $\alpha \in \mathcal{K}_i(\alpha)$. It follows from this and (6.6) that $\alpha \models \neg K_i \neg (s_i^\ell \succ_i s_i^k)$. Thus $\alpha \models s_i^k \wedge K_i(s_i^\ell \succeq_i s_i^k) \wedge \neg K_i \neg (s_i^\ell \succ_i s_i^k)$, so that axiom $\mathbf{SR}'$ is falsified at α. Q.E.D.

There appears to be an important difference between the results of Section 5 and those of this section, namely that, while Propositions 5.4 and 5.8 give a *local* result, Propositions 6.2 and 6.3 provide a *global* one. For example, Proposition 5.4 says that if *at a state* there is common belief of rationality, then the strategy profile played *at that state* belongs to S^∞, while its counterpart in this section, namely Proposition 6.2, says that the strategy profile played *at every state* belongs to S^∞. As a matter of fact, the results of Section 5 are also global in nature. Consider, for example, Proposition 5.4. Fix a model and a state α and suppose that $\alpha \models B_*(r_1 \wedge \cdots \wedge r_n)$. Since, for every formula φ, $B_*\varphi \to B_*B_*\varphi$ is a theorem of $\mathbf{KD45}_n^*$, it follows that $\alpha \models B_*B_*(r_1 \wedge \cdots \wedge r_n)$, that is, for every $\omega \in \mathcal{B}_*(\alpha)$, $\omega \models B_*(r_1 \wedge \cdots \wedge r_n)$. Thus, it follows from Proposition 5.4 that $\sigma(\omega) \in S^\infty$, for every $\omega \in \mathcal{B}_*(\alpha)$.[16] That is, if at a state there is common belief of rationality, then at that state, *as well as at all states reachable from it by the common belief relation* $\mathcal{B}_*$, it is true that the strategy profile played belongs to S^∞. This is essentially

[16] This fact was proved directly in the proof of Proposition 5.4.

a global result, since from the point of view of a state α, the "global" space is precisely the set $\mathcal{B}_*(\alpha)$.

Thus the only difference between the results of Section 5 and those of this section lies in the fact that Propositions 5.4 and 5.8 bring out the role of common belief by mimicking the informal argument that if player 1 is rational then she won't choose a strategy $s_1 \in D_1^0$ and if player 2 believes that player 1 is rational then he believes that $s_1 \notin D_1^0$ and therefore will not choose a strategy $s_2 \in D_2^1$, and if player 1 believes that player 2 believes that player 1 is rational, then player 1 believes that $s_2 \notin D_2^1$ and will, therefore, not choose a strategy $s_1 \in D_1^2$, and so on. Beliefs about beliefs about beliefs... are explicitly modeled through the common belief operator. In contrast, Propositions 6.2 and 6.3 do not make use of the common belief operator. However, the logic is essentially the same. In particular, common belief of rationality is generated by the axiom $\mathbf{WR}'$ (or $\mathbf{SR}'$) and the rule of necessitation: from $s_1^k \rightarrow \neg B_1(s_1^\ell \succ_1 s_1^k)$ we get, by Necessitation, that $B_1(s_1^k \rightarrow \neg B_1(s_1^\ell \succ_1 s_1^k)) \wedge B_2(s_1^k \rightarrow \neg B_1(s_1^\ell \succ_1 s_1^k))$ and thus, whatever is implied by $\mathbf{WR}'$ is believed by both players. Further iterations of the Necessitation rule yields beliefs about beliefs about beliefs... about the rationality of every player.

7 Related literature

As noted in the introduction, the iterated elimination of strictly dominated strategies as a solution concept for strategic-form games goes back to Bernheim [2] and Pearce [16] and has been further studied and characterized by a number of authors. From the point of view of this paper, the most important contribution in this area is due to Stalnaker [17], who put forward the novel proposal of characterizing solution concepts for games in terms of classes of models. Stalnaker carried out his analysis within the standard framework of von Neumann-Morgenstern payoffs and defined dominance in terms of mixed strategies. Furthermore his analysis was semantic rather than syntactic. Our approach differs from Stalnaker's in that we formulate rationality syntactically within an axiomatic system and provide characterization results in line with the notion of frame characterization in modal logic. Furthermore, we do this in a purely ordinal framework that does not require probabilistic beliefs and von Neumann-Morgenstern payoffs. However, our intellectual debt towards Stalnaker is clear. In particular, the IDIP algorithm (see Definition 3.4) is the adaptation to ordinal games of the algorithm he introduced in [17].

A syntactic epistemic analysis of iterated strict dominance was also proposed by de Bruin [9, p. 86]. However his approach is substantially different from ours. First of all, his analysis is explicitly carried out only for two-person games, while we allowed for any number of players. Secondly,

de Bruin assumes von Neumann Morgenstern payoffs and his definition of strict dominance involves domination by mixed strategies [9, p. 51], while we considered ordinal payoffs and defined dominance in terms of pure strategies only (see Definition 3.2). Thirdly, de Bruin restricts attention to knowledge (that is, in his axiom system he imposes the Truth Axiom $\mathbf{T}_i$ on individual beliefs: [9, p. 51]) and thus does not investigate the difference between the implications of common belief of rationality and those of common knowledge of rationality (hence in his analysis there is no counterpart to the difference highlighted in Propositions 5.4 and 5.8 of Section 5). More importantly, however, de Bruin introduces the notion of strict dominance *directly into the syntax* by using atomic propositions of the form $nsd_i(A_i, A_j)$ whose intended interpretation is "player i uses a pure strategy in A_i which is not strictly dominated by a mixed strategy over A_i given that player j plays a pure strategy in A_j". Furthermore, his definition of rationality *incorporates* the notion of strict dominance. De Bruin's definition of rationality [9, p. 86] consists of two parts: a basis step without knowledge, $r_i \rightarrow nsd_i(A_i, A_j)$, and an inductive step with knowledge: $(r_i \wedge K_i X_i \wedge K_i X_j) \rightarrow nsd_i(X_i, X_j)$. According to de Bruin the advantage of his two-part definition of rationality is that

> Drawing a line between a basis case without beliefs, and an inductive step with beliefs makes it possible to mimic every single round of elimination of the solution concept by a step in the hierarchy of common belief in rationality. This becomes highly explicit in the inductive character of the proof. [9, p. 100]

However, as pointed out at the end of the previous section, this mimicking of the elimination steps occurs also in the proof of Proposition 5.4 without the need for a two-part definition of rationality and, more importantly, without incorporating the notion of dominance in the syntax.

The disadvantage of de Bruin's approach is that one loses the distinction between syntax and semantics and the ability to link the two by means of frame characterization results. In our analysis, the notion of strict dominance is purely a semantic notion, which has no syntactic counterpart. On the other hand the definition of rationality is expressed syntactically and it is epistemically based, in that it evaluates a player's rationality by comparing her action with her beliefs about the desirability of alternative actions. The characterization results then establish a correspondence between the output of an algorithm (such as the iterated deletion of strictly dominated strategies) and common belief of an independently formulated notion of rationality.

Börgers [7] provides a characterization of pure-strategy dominance that differs from ours. Like us, Börgers assumes that only the ordinal rankings of the players are commonly known; however—unlike us—he also assumes that

	d	e
a	0,0	0,0
b	1,1	0,0
c	0,0	1,1

	d	e
a	0,0	0,0
b	1,1	0,0

FIGURE 8. Two games in which player 2 has strategies d and e, whereas player 1 has either three strategies (left) or two (right).

each player has a von Neumann-Morgenstern utility function on the set of outcomes, forms probabilistic beliefs about the opponents' strategy choices and chooses a pure strategy that maximizes her expected utility, given those beliefs. He thus asks the question: what pure-strategy profiles are consistent with common belief of rationality, where the latter is defined as expected utility maximization with respect to *some* von Neumann-Morgenstern utility function and *some* beliefs about the opponents' strategies? Börgers shows that a pure strategy is rational in this sense if and only if it is not dominated by another *pure* strategy. Thus there is no need to consider dominance by a mixed strategy. However, he shows that the relevant notion of dominance in this case is *not* strict dominance but the following stronger notion: a strategy $s_i \in S_i$ of player i is dominated if and only if, for every subset of strategy profiles $X_{-i} \subseteq S_{-i}$ of the players other than i, there exists a strategy $t_i \in S_i$ (which can vary with X_{-i}) that *weakly* dominates s_i relative to X_{-i}.[17] For example, in the game illustrated in Figure 8 (left), strategy a of player 1 is dominated (by b relative to $\{d, e\}$ and also relative to $\{d\}$ and by c relative to $\{e\}$). Thus in the corresponding model shown in Figure 9, at state α, while player 1 is rational according to our axiom **WR** (since no strategy is strictly dominated; indeed at state α there is common knowledge of rationality), she is not rational according to Börgers' definition.[18]

On the other hand, while stronger than the notion expressed by our axiom **WR**, Börgers' notion of rationality is *weaker* than our axiom **SR**, as can be seen in the game of Figure 8 (right). Here strategy a of player 1 is dominated by b relative to $\{d, e\}$ and also relative to $\{d\}$ but not relative to $\{e\}$. Thus a is a rational strategy according to Börgers' definition. On the other hand, in the model of Figure 9 (viewed now as a model for the game of Figure 8 (right)) player 1 is not rational at state α (where her choice is

[17] We say that t_i weakly dominates s_i relative to X_{-i} if (1) $z(t_i, x_{-i}) \succeq_i z(s_i, x_{-i})$, for all $x_{-i} \in X_{-i}$, and (2) there exists an $\hat{x}_{-i} \in X_{-i}$ such that $z(t_i, \hat{x}_{-i}) \succ_i z(s_i, \hat{x}_{-i})$.

[18] The reason for this is as follows: if player 1 assigns positive probability to both α and β, then she would get a higher expected utility by switching to strategy b. The same is true if she assigns probability 1 to α. On the other hand, if she assigns probability 1 to β then she can increase her utility by switching to c.

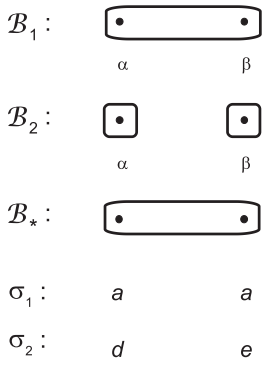


FIGURE 9. A model for the games of Figure 8.

a) according to the notion of rationality expressed by axiom **SR** (indeed, in this game, $T^\infty = \{(a, e), (b, d)\}$ and thus $(a, d) \notin T^\infty$).

8 Conclusion

We have examined the implications of common belief and common knowledge of two, rather weak, notions of rationality. Most of the literature on the epistemic foundations of game theory have dealt with the Bayesian approach, which identifies rationality with expected payoff maximization, given probabilistic beliefs (for surveys of this literature see [1] and [11]). Our focus has been on strategic-form games with ordinal payoffs and non-probabilistic beliefs. While most of the literature has been developed within the semantic approach, we have used a syntactic framework and expressed rationality in terms of syntactic axioms. We showed that the first, weaker, axiom of rationality characterizes the iterated deletion of strictly dominated strategies, while the stronger axiom characterizes the pure-strategy version of the algorithm introduced by Stalnaker [17].

The two notions of rationality used in this paper can, of course, be used also in the subclass of games with von Neumann-Morgenstern payoffs and the results would be the same. Furthermore, the standard notion of Bayesian rationality as expected payoff maximization is stronger than (that is, implies) both notions of rationality considered in this paper. Thus our results apply also to Bayesian rationality.[19]

We have provided two versions of our characterization results. The first (Propositions 5.4 and 5.8), which comes closer to the previous game-theoretic literature, is based on an explicit account of the role of common

[19] In the sense that whatever is incompatible with our notion of rationality is also incompatible with the stronger notion of Bayesian rationality.

belief of rationality and thus requires a syntax that contains atomic propositions that are interpreted as "player i is rational". The second characterization (Propositions 6.2 and 6.3) is closer to the modal logic literature, where axioms are characterized in terms of properties of frames. However, we argued that the two characterizations are essentially identical.

We have restricted attention to strategic-form games. In future work we intend to extend this qualitative (that is, non probabilistic) analysis to extensive-form games with perfect information and the notion of backward induction.

References

[1] P. Battigalli & G. Bonanno. Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics*, 53(2):149–225, 1999.

[2] D. Bernheim. Rationalizable strategic behavior. *Econometrica*, 52(4):1002–1028, 1984.

[3] P. Blackburn, M. de Rijke & Y. Venema. *Modal Logic*. Cambridge University Press, 2001.

[4] G. Bonanno. On the logic of common belief. *Mathematical Logic Quarterly*, 42(1):305–311, 1996.

[5] G. Bonanno & K. Nehring. On Stalnaker's notion of strong rationalizability and Nash equilibrium in perfect information games. *Theory and Decision*, 45(3):291–295, 1998.

[6] G. Bonanno & K. Nehring. Common belief with the logic of individual belief. *Mathematical Logic Quarterly*, 46(1):49–52, 2000.

[7] T. Börgers. Pure strategy dominance. *Econometrica*, 61(2):423–430, 1993.

[8] A. Brandenburger & E. Dekel. Rationalizability and correlated equilibria. *Econometrica*, 55(6):1391–1402, 1987.

[9] B. de Bruin. *Explaining Games: On the Logic of Game Theoretic Explanations*. Ph.D. thesis, University of Amsterdam, 2004. *ILLC Publications* DS-2004-03.

[10] B. Chellas. *Modal Logic: An Introduction*. Cambridge University Press, 1984.

[11] E. Dekel & F. Gul. Rationality and knowledge in game theory. In D. Kreps & K. Wallis, eds., *Advances in economics and econometrics*, pp. 87–172. Cambridge University Press, 1997.

[12] R. Fagin, J. Halpern, Y. Moses & M. Vardi. *Reasoning about Knowledge*. MIT Press, 1995.

[13] S. Kripke. A semantical analysis of modal logic I: normal propositional calculi. *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik*, 9:67–96, 1963.

[14] L. Lismont. La connaissance commune en logique modale. *Mathematical Logic Quarterly*, 39(1):115–130, 1993.

[15] L. Lismont & P. Mongin. On the logic of common belief and common knowledge. *Theory and Decision*, 37(1):75–106, 1994.

[16] D. Pearce. Rationalizable strategic behavior and the problem of perfection. *Econometrica*, 52(4):1029–1050, 1984.

[17] R. Stalnaker. On the evaluation of solution concepts. *Theory and Decision*, 37(1):49–74, 1994.

[18] T. Tan & S. Werlang. The Bayesian foundation of solution concepts of games. *Journal of Economic Theory*, 45(2):370–391, 1988.

Semantic Results for Ontic and Epistemic Change

Hans van Ditmarsch[1,2]
Barteld Kooi[3]

[1] Department of Computer Science
University of Otago
PO Box 56
Dunedin 9054, New Zealand

[2] Institut de Recherche en Informatique de Toulouse
Université Paul Sabatier
118 Route de Narbonne
31062 Toulouse Cedex 9, France

[3] Faculteit Wijsbegeerte
Rijksuniversiteit Groningen
Oude Boteringestraat 52
9712 GL Groningen, The Netherlands

hans@cs.otago.ac.nz, B.P.Kooi@rug.nl

Abstract

We present an epistemic logic incorporating dynamic operators to describe information changing events. Such events include epistemic changes, where agents become more informed about the non-changing state of the world, and ontic changes, wherein the world changes. The events are executed in information states that are modelled as pointed Kripke models. Our contribution consists of three semantic results. (i) Every consistent formula can be made true in every information state by the execution of an event. (ii) Every event corresponds to an event with assignments to true and false only. (iii) Every event corresponds to a sequence of events with assignments of a single atom only. We apply the logic to model dynamics in a multi-agent setting involving card players.

1 Introduction

In dynamic epistemic logics [32, 23, 9, 4, 17] one does not merely describe the static (knowledge and) beliefs of agents but also dynamic features: how does belief change as a result of events taking place. The main focus of such logics has been change of *only* belief, whereas the facts describing the world remain the same. Change of belief is known as *epistemic* change. One can also model change of facts, and the resulting consequences of such factual

Giacomo Bonanno, Wiebe van der Hoek, Michael Wooldridge (eds.). *Logic and the Foundations of Game and Decision Theory (LOFT 7).* Texts in Logic and Games 3, Amsterdam

changes for the beliefs of the agents. Change of facts is also known as *ontic* change (change of the real world, so to speak).[1] In this contribution we use 'event' to denote *any* sort of information change, both epistemic and ontic. Let us begin by a simple example involving various events.

Example 1.1. Given are two players Anne and Bill. Anne shakes a cup containing a single coin and deposits the cup upside down on the table (there are no opportunities for cheating). Heads or tails? Initially, we have a situation wherein both Anne (a) and Bill (b) are uncertain about the truth of that proposition. A player may observe whether the coin is heads or tails, and/or flip the coin, and with or without the other player noticing that. Four example events are as follows.

1. Anne lifts the cup and looks at the coin. Bill observes this but is not able to see the coin. All the previous is common knowledge to Anne and Bill.

2. Anne lifts the cup and looks at the coin without Bill noticing that.

3. Anne lifts the cup, looks at the coin, and ensures that it is tails (by some sleight of hand). Bill observes Anne looking but is not able to see the coin, and he considers it possible that Anne has flipped the coin to tails (and this is common knowledge).

4. Bill flips the coin (without seeing it). Anne considers that possible (and this is common knowledge).

Events 1, 3, and 4 are all public in the sense that the actual event is considered possible by both agents, and that both agents know that, and know that they know that, etc.; whereas event 2 is private: Bill is unaware of the event; the event is private to Anne. Events 3 and 4 involve ontic change, whereas events 1 and 2 only involve epistemic change. Flipping a coin is ontic change: the value of the atomic proposition 'the coin is heads' changes from false to true, or from true to false, because of that. But in the case of events 1 and 2 that value, whether true or false, remains unchanged. What changes instead, is how informed the agents are about that value, or about how informed the other agent is. In 1 and 2, Anne still learns whether the coin is heads or tails. In 1, Bill 'only' learns that Anne has learnt whether the coin is heads or tails: he has not gained factual information at all. In Example 2.5, later, we will formalize these descriptions.

[1] In the areas known as 'artificial intelligence' and 'belief revision', epistemic and ontic change are called, respectively, *belief revision* [1] and *belief update* [27]. We will not use that terminology.

Various logics have been proposed to model such events. A well-known setting is that of interpreted systems by Fagin et al. [21]. Each agent has a local state; the local states of all agents together with a state of the environment form a global state; belief of an agent is modelled as uncertainty to distinguish between global states wherein the agent has the same local state, and change of belief is modelled as a transition from one global state to another one, i.e., as a next step in a run through the system. In an interpreted system the treatment of epistemic and ontic change is similar—either way it is just a next step in a run, and how the valuation between different points changes is not essential to define or describe the transition. There is a long tradition in such research [30, 21].

The shorter history of dynamic epistemic logic started by focusing on epistemic change [32, 23, 9, 4, 17]. In that community, how to model ontic change was first mentioned by Baltag, Moss, and Solecki as a possible extension to their *action model logic* for epistemic change [5]. More detailed proposals for ontic change are far more recent [20, 16, 15, 10, 28, 33, 25, 26]. The literature will be discussed in more detail in Section 5.

Section 2 contains logical preliminaries, including detailed examples. Section 3 contains the semantic results that we have achieved for the logic. This is our original contribution to the area. These results are that: *(i)* for all finite models and for all consistent formulas we can construct an event that 'realizes' the formula, i.e., the formula becomes true after execution of the event; that: *(ii)* every event (with assignments of form $p := \varphi$, for φ in the language) corresponds to an event with assignments to true and false only; and also that: *(iii)* every event corresponds to a sequence of events with assignments for a single atom only. Section 4 applies the logic to model the dynamics of card players. Section 5 discusses related work in detail.

2 A logic of ontic and epistemic change

We separately introduce the logical language, the relevant structures, and the semantics of the language on those structures. The syntax and semantics *appear* to overlap: updates are structures that come with preconditions that are formulas. In fact it is properly covered by the double induction used in the language definition, as explained after Definition 2.4 below. For a detailed treatment of the logic without ontic change, and many examples, we recommend [17]; for more examples of the logic involving ontic change, see [15, 16, 20].

2.1 Language

We use the style of notation from propositional dynamic logic (PDL) for modal operators which is also used in [10].

Definition 2.1 (Language)**.** Let a finite set of agents A and a countable set of propositional variables P be given. The language $\mathscr{L}$ is given by the following BNF:

$$\begin{array}{lcl} \varphi & ::= & p \mid \neg\varphi \mid \varphi \wedge \varphi \mid [\alpha]\varphi \\ \alpha & ::= & a \mid B^* \mid (\mathsf{U}, \mathsf{e}) \end{array}$$

where $p \in P$, $a \in A$, $B \subseteq A$ (the dynamic operator $[B^*]$ is associated with 'common knowledge among the agents in B'), and where (U, e) is an *update* as (simultaneously) defined below.

We use the usual abbreviations, and conventions for deleting parentheses. In particular, $[B]\varphi$ stands for $\bigwedge_{a \in B}[a]\varphi$, and (the diamond form) $\langle\alpha\rangle\varphi$ is equivalent to $\neg[\alpha]\neg\varphi$. Non-deterministic updates are introduced by abbreviation: $[(\mathsf{U}, \mathsf{e}) \cup (\mathsf{U}', \mathsf{f})]\varphi$ is by definition $[\mathsf{U}, \mathsf{e}]\varphi \wedge [\mathsf{U}', \mathsf{f}]\varphi$.

2.2 Structures

Epistemic model. The models which adequately present an information state in a multi-agent environment are Kripke models from epistemic logic. The set of states together with the accessibility relations represent the information the agents have. If one state s has access to another state t for an agent a, this means that, if the actual situation is s, then according to a's information it is possible that t is the actual situation.

Definition 2.2 (Epistemic model)**.** Let a finite set of agents A and a countable set of propositional variables P be given. An epistemic model is a triple $M = (S, R, V)$ such that

- *domain* S is a non-empty set of possible states,
- $R : A \to \wp(S \times S)$ assigns an *accessibility relation* to each agent a,
- $V : P \to \wp(S)$ assigns a set of states to each propositional variable; this is the *valuation* of that variable.

A pair (S, R) is called an *epistemic frame*. A pair (M, s), with $s \in S$, is called an *epistemic state*.

A well-known notion of sameness of epistemic models is 'bisimulation'. Several of our results produce models that are bisimilar: they correspond in the sense that even when not identical (isomorphic), they still cannot be distinguished in the language.

Definition 2.3 (Bisimulation)**.** Let two models $M = (S, R, V)$ and $M' = (S', R', V')$ be given. A non-empty relation $\mathfrak{R} \subseteq S \times S'$ is a bisimulation iff for all $s \in S$ and $s' \in S'$ with $(s, s') \in \mathfrak{R}$:

atoms for all $p \in P$: $s \in V(p)$ iff $s' \in V'(p)$;

forth for all $a \in A$ and all $t \in S$: if $(s,t) \in R(a)$, then there is a $t' \in S'$ such that $(s',t') \in R'(a)$ and $(t,t') \in \mathfrak{R}$;

back for all $a \in A$ and all $t' \in S'$: if $(s',t') \in R'(a)$, then there is a $t \in S$ such that $(s,t) \in R(a)$ and $(t,t') \in \mathfrak{R}$.

We write $(M,s) \underline{\leftrightarrow} (M',s')$, iff there is a bisimulation between M and M' linking s and s', and we then call (M,s) and (M',s') bisimilar. A model such that all bisimilar states are identical is called a *bisimulation contraction* (also known as a *strongly extensional model*).

Update model. An epistemic model represents the information of the agents. *Information change* is modelled as changes of such a model. There are three variables. One can change the set of states, the accessibility relations and the valuation. It may be difficult to find the exact change of these variables that matches a certain description of an information changing event. It is often easier to think of such an event separately. One can model an information changing event in the same way as an information state, namely as some kind of Kripke model: there are various possible events, which the agents may not be able to distinguish. This is the domain of the model. Rather than a valuation, a *precondition* captures the conditions under which such events may occur, and *postconditions* also determine what epistemic models may evolve into. Such a Kripke model for events is called an *update model*, which were first studied by Baltag, Moss and Solecki, and extended with simultaneous substitutions by van Eijck [5, 20].[2] Here we use van Eijck's definition.

Definition 2.4 (Update model)**.** An *update model* for a finite set of agents A and a language $\mathscr{L}$ is a quadruple $\mathsf{U} = (\mathsf{E}, \mathsf{R}, \mathsf{pre}, \mathsf{post})$ where

- *domain* E is a finite non-empty set of events,
- $\mathsf{R} : A \to \wp(\mathsf{E} \times \mathsf{E})$ assigns an *accessibility relation* to each agent,
- $\mathsf{pre} : \mathsf{E} \to \mathscr{L}$ assigns to each event a *precondition*,
- $\mathsf{post} : \mathsf{E} \to (P \to \mathscr{L})$ assigns to each event a *postcondition* for each atom. Each $\mathsf{post}(\mathsf{e})$ is required to be only finitely different from the identity id; the finite difference is called its *domain* $\mathrm{dom}(\mathsf{post}(\mathsf{e}))$.

A pair (U, e) with a distinguished actual event $\mathsf{e} \in \mathsf{E}$ is called an *update*. A pair (U, E) with $\mathsf{E}' \subseteq \mathsf{E}$ and $|\mathsf{E}'| > 1$ is a *multi-pointed update*, first introduced in [19]. The event e with $\mathsf{pre}(\mathsf{e}) = \top$ and $\mathsf{post}(\mathsf{e}) = \mathsf{id}$ we name

[2] In the literature update models are also called action models. Here we follow [10] and call them update models, since no agency seems to be involved.

skip. An update with a singleton domain, accessible to all agents, and precondition $\top$, is a *public assignment*. An update with a singleton domain, accessible to all agents, and identity postcondition, is a *public announcement*.

Instead of

$$\mathsf{pre}(\mathsf{e}) = \varphi \text{ and } \mathsf{post}(\mathsf{e})(p_1) = \psi_1, \ldots, \text{ and } \mathsf{post}(\mathsf{e})(p_n) = \psi_n$$

we also write[3]

$$\text{for event } \mathsf{e}\text{: if } \varphi, \text{ then } p_1 := \psi_1, \ldots, \text{ and } p_n := \psi_n$$

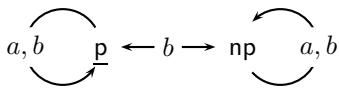
The event skip stands for: nothing happens except a tick of the clock.

To see an update as part of the language we observe that: an update (U, e) is an inductive construct of type α that is built the frame underlying U (we can assume a set enumerating such frames) and from simpler constructs of type φ, namely the preconditions and postconditions for the events of which the update consists. This means that there should be a finite number of preconditions and a finite number of postconditions only, otherwise the update would be an infinitary construct. A finite number of preconditions is guaranteed by restricting ourselves in the language to *finite* update models. A finite number of postconditions is guaranteed by (as well) restricting ourselves to *finite* domain for postconditions. This situation is similar to the case of automata-PDL [24, Chapter 10, Section 3].

If in case of nondeterministic updates the underlying models are the same, we can also see this as executing a multi-pointed update. For example, $(\mathsf{U}, \mathsf{e}) \cup (\mathsf{U}, \mathsf{f})$ can be equated with $(\mathsf{U}, \{\mathsf{e}, \mathsf{f}\})$.

Example 2.5. Consider again the scenario of Example 1.1 on page 88. Let atomic proposition p stand for 'the coin lands heads'. The initial information state is represented by a two-state epistemic model with domain $\{1, 0\}$, with universal access for a and b, and with $V(p) = \{1\}$. We further assume that the actual state is 1. This epistemic state is depicted in the top-left corner of Figure 1. The events in Example 1.1 can be visualized as the following updates. The actual state is underlined.

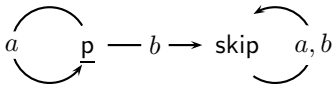
1. Anne lifts the cup and looks at the coin. Bill observes this but is not able to see the coin. All the previous is common knowledge to Anne and Bill.



[3] The notation is reminiscent of that for a *knowledge-based program* in the interpreted systems tradition. We discuss the correspondence in Section 5.

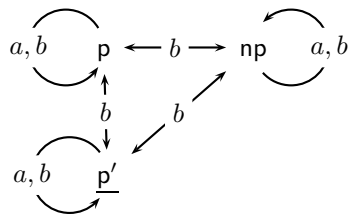
Here, $\mathsf{pre}(\mathsf{p}) = p$, $\mathsf{post}(\mathsf{p}) = \mathsf{id}$, $\mathsf{pre}(\mathsf{np}) = \neg p$, $\mathsf{post}(\mathsf{np}) = \mathsf{id}$. The update model consists of two events. The event p corresponds to Anne seeing heads, and the event np to Anne seeing tails; Anne is aware of that: thus the reflexive arrows (identity relation). Bill cannot distinguish them from one another: thus the universal relation. The aspect of common knowledge, or common awareness, is also present in this dynamic structure: the reflexive arrow for Anne also encodes that Anne knows that she lifts the cup and that Bill observes that; similarly for Bill, and for iterations of either awareness.

2. Anne lifts the cup and looks at the coin without Bill noticing that.



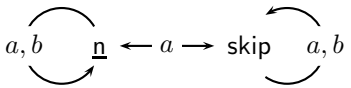
Event p is as in the previous item and skip is as above. In this update, there is no common awareness of what is going on: Anne observes heads knowing that Bill is unaware of that, whereas Bill does not consider the actual event; the b-arrow points to the other event only.

3. Anne lifts the cup, looks at the coin, and ensures that it is tails (by some sleight of hand). Bill observes Anne looking but is not able to see the coin, and he considers it possible that Anne has flipped the coin to tails (and this is common knowledge).



Events p and np are as before, whereas $\mathsf{pre}(\mathsf{p}') = \top$, $\mathsf{post}(\mathsf{p}')(p) = \bot$. The event p' may take place both when the coin is heads and when the coin is tails, in the first case atom p is set to false (tails), and in the second it remains false.

4. Bill flips the coin (without seeing it). Anne considers that possible (and this is common knowledge).



Here, $\mathsf{pre}(\mathsf{n}) = \top$, $\mathsf{post}(\mathsf{n})(p) = \neg p$, and skip is as before. For models where all accessibility relations are equivalence relations we will also use a simplified visualization that merely links states in the same equivalence class. E.g., this final event is also depicted as:

$$\underline{\mathsf{n}} \;\text{—}\; a \;\text{—}\; \mathsf{skip}$$

2.3 Semantics

The semantics of this language is standard for epistemic logic and based on the product construction for the execution of update models from the previous section. Below, $R(B)^*$ is the transitive and reflexive closure of the union of all accessibility relations $R(a)$ for agents $a \in B$. Definitions 2.6 and 2.7 are supposed to be defined simultaneously.

Definition 2.6 (Semantics)**.** Let a model (M, s) with $M = (S, R, V)$ be given. Let $a \in A$, $B \subseteq A$, and $\varphi, \psi \in \mathscr{L}$.

$$\begin{array}{lll}
(M,s) \models p & \text{iff} & s \in V(p) \\
(M,s) \models \neg\varphi & \text{iff} & (M,s) \not\models \varphi \\
(M,s) \models \varphi \wedge \psi & \text{iff} & (M,s) \models \varphi \text{ and } (M,s) \models \psi \\
(M,s) \models [a]\varphi & \text{iff} & (M,t) \models \varphi \text{ for all } t \text{ such that } (s,t) \in R(a) \\
(M,s) \models [B^*]\varphi & \text{iff} & (M,t) \models \varphi \text{ for all } t \text{ such that } (s,t) \in R(B)^* \\
(M,s) \models [\mathsf{U},\mathsf{e}]\varphi & \text{iff} & (M,s) \models \mathsf{pre}(\mathsf{e}) \text{ implies } (M \otimes \mathsf{U}, (s,\mathsf{e})) \models \varphi
\end{array}$$

We now define the effect of an update on an epistemic state—Figure 1 gives an example of such update execution.

Definition 2.7 (Execution)**.** Given are an epistemic model $M = (S, R, V)$, a state $s \in S$, an update model $\mathsf{U} = (\mathsf{E}, \mathsf{R}, \mathsf{pre}, \mathsf{post})$, and an event $\mathsf{e} \in \mathsf{E}$ with $(M, s) \models \mathsf{pre}(\mathsf{e})$. The result of executing (U, e) in (M, s) is the model $(M \otimes \mathsf{U}, (s, \mathsf{e})) = ((S', R', V'), (s, \mathsf{e}))$ where

- $S' = \{(t, \mathsf{f}) \mid (M, t) \models \mathsf{pre}(\mathsf{f})\}$,
- $R'(a) = \{((t, \mathsf{f}), (u, \mathsf{g})) \mid (t, \mathsf{f}), (u, \mathsf{g}) \in S' \text{ and } (t, u) \in R(a) \text{ and } (\mathsf{f}, \mathsf{g}) \in \mathsf{R}(a)\}$,
- $V'(p) = \{(t, \mathsf{f}) \mid (M, t) \models \mathsf{post}(\mathsf{f})(p)\}$.

Definition 2.8 (Composition of update models)**.** Let update models $\mathsf{U} = (\mathsf{E}, \mathsf{R}, \mathsf{pre}, \mathsf{post})$ and $\mathsf{U}' = (\mathsf{E}', \mathsf{R}', \mathsf{pre}', \mathsf{post}')$ and events $\mathsf{e} \in \mathsf{E}$ and $\mathsf{e}' \in \mathsf{E}'$ be given. The composition (U, e) ; $(\mathsf{U}', \mathsf{e}')$ of these update models is $(\mathsf{U}'', \mathsf{e}'')$ where $\mathsf{U}'' = (\mathsf{E}'', \mathsf{R}'', \mathsf{pre}'', \mathsf{post}'')$ is defined as

- $\mathsf{E}'' = \mathsf{E} \times \mathsf{E}'$,

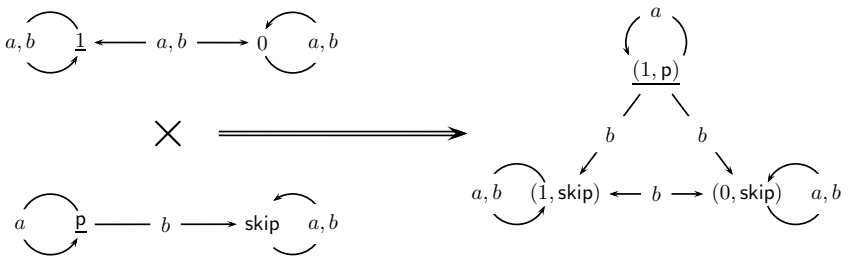


FIGURE 1. In an epistemic state where Anne and Bill are uncertain about the truth of p (heads or tails), and wherein p is true, Anne looks at the coin without Bill noticing it. The actual states and events are underlined.

- $\mathsf{R}''(a) = \{((\mathsf{f},\mathsf{f}'),(\mathsf{g},\mathsf{g}')) \mid (\mathsf{f},\mathsf{g}) \in \mathsf{R}(a) \text{ and } (\mathsf{f}',\mathsf{g}') \in \mathsf{R}'(a)\}$,
- $\mathsf{pre}''(\mathsf{f},\mathsf{f}') = \mathsf{pre}(\mathsf{f}) \wedge [\mathsf{U},\mathsf{f}]\mathsf{pre}'(\mathsf{f}')$,
- $\mathrm{dom}(\mathsf{post}''(\mathsf{f},\mathsf{f}')) = \mathrm{dom}(\mathsf{post}(\mathsf{f})) \cup \mathrm{dom}(\mathsf{post}'(\mathsf{f}'))$ and if $p \in \mathrm{dom}(\mathsf{post}''(\mathsf{f},\mathsf{f}'))$, then

$$\mathsf{post}''(\mathsf{f},\mathsf{f}')(p) = \begin{cases} \mathsf{post}(\mathsf{f})(p) & \text{if } p \notin \mathrm{dom}(\mathsf{post}'(\mathsf{f}')), \\ [\mathsf{U},\mathsf{f}]\mathsf{post}'(\mathsf{f}')(p) & \text{otherwise.} \end{cases}$$

The reason for $[\mathsf{U},\mathsf{f}]\mathsf{post}'(\mathsf{f}')(p)$ in the final clause will become clear from the proof detail shown for Proposition 2.9.

Proposition 2.9. $\models [\mathsf{U},\mathsf{e}][\mathsf{U}',\mathsf{e}']\varphi \leftrightarrow [(\mathsf{U},\mathsf{e}) \;;\; (\mathsf{U}',\mathsf{e}')]\varphi$

Proof. Let (M,t) be arbitrary. To show that $(M,t) \models [(\mathsf{U},\mathsf{e}) \;;\; (\mathsf{U}',\mathsf{e}')]\varphi$ if and only if $(M,t) \models [\mathsf{U},\mathsf{e}][\mathsf{U}',\mathsf{e}']\varphi$, it suffices to show that $M \otimes (\mathsf{U} \;;\; \mathsf{U}')$ is isomorphic to $(M \otimes \mathsf{U}) \otimes \mathsf{U}'$. A detailed proof (for purely epistemic updates) is found in [17]. The postconditions (only) play a part in the proof that the valuations correspond:

For the valuation of facts p in the domain of post'' we distinguish the cases ($p \in \mathrm{dom}(\mathsf{post}(\mathsf{e}))$ but $p \notin \mathrm{dom}(\mathsf{post}'(\mathsf{e}'))$) *(i)*, and otherwise *(ii)*. The valuation of a i-atom in a triple $(t,(\mathsf{e},\mathsf{e}'))$ is $\mathsf{post}(\mathsf{e})(p)$ according to the definition of updates composition; and the valuation of a ii-atom is $[\mathsf{U},\mathsf{e}]\mathsf{post}'(\mathsf{e}')(p)$. Consider the corresponding triple $((t,\mathsf{e}),\mathsf{e}')$. The valuation of an i-atom in (t,e) is $\mathsf{post}(\mathsf{e})(p)$, and because p does not occur in $\mathrm{dom}(\mathsf{post}'(\mathsf{e}'))$ its value in the triple $((t,\mathsf{e}),\mathsf{e}')$ will remain the same. For a ii-atom, its final value will be determined by evaluating $\mathsf{post}'(\mathsf{e}')(p)$ in $((M \otimes \mathsf{U}),(t,\mathsf{e}))$. This corresponds to evaluating $[\mathsf{U},\mathsf{e}]\mathsf{post}'(\mathsf{e}')(p)$ in (M,t). Q.E.D.

2.4 Proof system

A proof system **UM** for the logic is given in Table 1. The proof system is a lot like the proof system for the logic of epistemic actions (i.e., for the logic *without* postconditions to model valuation change) in [5]. There are two differences that makes it worthwhile to present this system. The axiom 'atomic permanence' in [5]—$[\mathsf{U},\mathsf{e}]p \leftrightarrow (\mathsf{pre}(\mathsf{e}) \rightarrow p)$—is now instead an axiom expressing when atoms are *not* permanent, namely how the value of an atom can change, according to the postcondition for that atom:

$$[\mathsf{U},\mathsf{e}]p \leftrightarrow (\mathsf{pre}(\mathsf{e}) \rightarrow \mathsf{post}(\mathsf{e})(p)) \qquad \text{update and atoms}$$

The second difference is not apparent from Table 1. The axiom

$$[\mathsf{U},\mathsf{e}][\mathsf{U}',\mathsf{e}']\varphi \leftrightarrow [(\mathsf{U},\mathsf{e}) \;;\; (\mathsf{U}',\mathsf{e}')]\varphi \qquad \text{update composition}$$

also occurs in [5]. But Definition 2.8 to compute that composition is in our case a more complex construction than the composition of update models with only preconditions, because it also involves resetting the postconditions. We find it remarkable that these are the only differences: the interaction between *postconditions* for an atom and the logical operators, *only* occurs in the axiom where that atom is mentioned, or implicitly, whereas the interaction between *preconditions* and the logical operators appears in several axioms and rules.

The proof system is sound and complete. The soundness of the 'update and atoms' axiom is evident. The soundness of the 'update composition' axiom was established in Proposition 2.9. We proved completeness of the logic as a modification of the completeness proof for the logic without ontic change—action model logic—as found in [17], which in turn is a simplified version of the original proof for that logic as found in [5]. We do not consider the modified proof of sufficient original interest to report on in detail.

3 Semantic results

We now present some semantic peculiarities of the logic. These we deem our contribution to the area. The results help to relate different approaches combining ontic and epistemic change (see Section 5). The various 'normal forms' for update models that we define are also intended to facilitate future tool development. Finally, they are relevant when modelling AGM belief revision [1] in a dynamic epistemic setting.

3.1 Arbitrary belief change

Let (M, s) and (M', s') be arbitrary finite epistemic states for the same set of atoms and agents, with $M = (S, R, V)$ and $M' = (S', R', V')$. Surprisingly enough, there is almost always an update transforming the former into the

All instantiations of propositional tautologies	
$[\alpha](\varphi \to \psi) \to ([\alpha]\varphi \to [\alpha]\psi)$	distribution
From φ and $\varphi \to \psi$, infer ψ	modus ponens
From φ, infer $[\alpha]\varphi$	necessitation
$[\mathsf{U},\mathsf{e}]p \leftrightarrow (\mathsf{pre}(\mathsf{e}) \to \mathsf{post}(\mathsf{e})(p))$	update and atoms
$[\mathsf{U},\mathsf{e}]\neg\varphi \leftrightarrow (\mathsf{pre}(\mathsf{e}) \to \neg[\mathsf{U},\mathsf{e}]\varphi)$	update and negation
$[\mathsf{U},\mathsf{e}](\varphi \wedge \psi) \leftrightarrow ([\mathsf{U},\mathsf{e}]\varphi \wedge [\mathsf{U},\mathsf{e}]\psi)$	update and conjunction
$[\mathsf{U},\mathsf{e}][a]\varphi \leftrightarrow (\mathsf{pre}(\mathsf{e}) \to \bigwedge_{(\mathsf{e},\mathsf{f})\in \mathsf{R}(a)}[a][\mathsf{U},\mathsf{f}]\varphi)$	update and knowledge
$[\mathsf{U},\mathsf{e}][\mathsf{U}',\mathsf{e}']\varphi \leftrightarrow [(\mathsf{U},\mathsf{e});(\mathsf{U}',\mathsf{e}')]\varphi$	update composition
$[B^*]\varphi \to (\varphi \wedge [B][B^*]\varphi)$	mix
$[B^*](\varphi \to [B]\varphi) \to (\varphi \to [B^*]\varphi)$	induction axiom
Let (U,e) be an update model and let a set of formulas χ_f for every f such that $(\mathsf{e},\mathsf{f}) \in \mathsf{R}(B)^*$ be given. From $\chi_\mathsf{f} \to [\mathsf{U},\mathsf{f}]\varphi$ and $(\chi_\mathsf{f} \wedge \mathsf{pre}(\mathsf{f})) \to [a]\chi_\mathsf{g}$ for every $\mathsf{f} \in \mathsf{E}$ such that $(\mathsf{e},\mathsf{f}) \in \mathsf{R}(B)^*$, $a \in B$ and $(\mathsf{f},\mathsf{g}) \in \mathsf{R}(a)$, infer $\chi_\mathsf{e} \to [\mathsf{U},\mathsf{e}][B^*]\varphi$.	updates and common knowledge

TABLE 1. The proof system **UM**.

latter. There are two restrictions. Both restrictions are technical and not conceptual. Firstly, for the set of agents with non-empty access in M' there must be a submodel of M containing actual state s that is serial for those agents. In other words, if an agent initially has empty access and therefore believes everything ('is crazy') you cannot change his beliefs, but otherwise you can. This seems reasonable. Secondly, models M and M' should only differ in the value of a finite number of atoms; more precisely, if we define that

> an atom is *relevant* in a model iff its valuation is neither empty nor the entire domain,

then the requirement is that only a finite number of atoms is relevant in $M \cup M'$. This is required, because we can only change the value of a finite number of atoms in the postconditions. This also seems reasonable: as both models are finite, the agents can only be uncertain about the value of a *finite* number of atoms (in the combined models M and M'); in other words, they are 'not interested' in the value of the remaining atoms.

For expository purposes we initially assume that all agents consider the actual state s of M a possibility (as in all S5 models, such as Kripke models representing interpreted systems), thus satisfying the first of the two restrictions above: the serial submodel required is then the singleton model

consisting of s, accessible to all agents. The update transforming (M, s) into (M', s') can be seen as the composition of two intuitively more appealing updates. That will make clear how we can also describe the required update in one stroke.

In the first step we get rid of the structure of M. As the epistemic state (M, s) is finite, it has a characteristic formula $\delta_{(M,s)}$ [6, 8].[4] We let the agents publicly learn that characteristic formula. This event is represented by the singleton update (U, e) defined as

$$((\{\mathsf{e}\}, \mathsf{R}, \mathsf{pre}, \mathsf{post}), \mathsf{e}) \text{ with } \begin{array}{l} \mathsf{pre}(\mathsf{e}) = \delta_{(M,s)}, \\ \text{for all } a : (\mathsf{e}, \mathsf{e}) \in \mathsf{R}(a), \\ \text{and } \mathsf{post}(\mathsf{e}) = \mathsf{id} \end{array}$$

In other words, the structure of the current epistemic state is being publicly announced. The resulting epistemic state is, of course, also singleton, or bisimilar to a singleton epistemic state, as $\delta_{(M,s)}$ holds in all states in M bisimilar to s. Without loss of generality assume that it is singleton. Its domain is $\{(s, \mathsf{e})\}$. This pair (s, e) is accessible to itself because for all agents, $(s, s) \in R(a)$ (all agents consider the actual state s a possibility), and $(\mathsf{e}, \mathsf{e}) \in \mathsf{R}(a)$. The valuation of propositional variables in this intermediate state are those of state s in M. What the value is does not matter: we will not use that valuation.

Now proceed to the second step. In the epistemic state wherein the agents have common knowledge of the facts in s, the agents learn the structure of the resulting epistemic state $M' = (S', R', V')$ and their part in it by executing update (U', s') defined as

$$((S', R', \mathsf{pre}', \mathsf{post}'), s') \text{ with } \begin{array}{l} \text{for all } t' \in S' : \mathsf{pre}'(t') = \top \text{ and} \\ \text{for relevant } p : \mathsf{post}'(t')(p) = \top \text{ iff } t' \in V'(p) \end{array}$$

Note that the domain S' and the accessibility relation R' of U' are precisely those of M', the resulting final epistemic model. The postcondition post' is well-defined, as only a finite number of atoms (the relevant atoms) is considered. Because we execute this update in a singleton model with public access, and because it is executable for every event t', the resulting epistemic state has the same structure as the update: it returns S' and R' again. The postcondition delivers the required valuation of atoms in the final model: for each *event* t' in U' and for all relevant atoms p, p become true in t' if p is true in *state* t' in M' ($\mathsf{post}(t')(p) = \top$), else p becomes false. The value of irrelevant atoms remains the same.

[4] A characteristic formula φ for a state (M, s) satisfies that for all ψ, $(M, s) \models \varphi$ iff $\varphi \models \psi$. In fact, for the construction we only need formulas that can distinguish all states in the domain from one another, modulo bisimilarity. Characteristic formulas satisfy that requirement.

We combine these two updates into one by requiring the precondition of the first and the postcondition of the second. Consider U'_r that is exactly as U' except that in all events t' in its domain the precondition is not $\top$ but $\delta_{(M,s)}$: the characteristic formula of *the point* s of (M, s). Update (U'_r, s') does the job: epistemic state $(M \otimes \mathsf{U}'_r, (s, s'))$ is isomorphic to (M', s'). This will be Corollary 3.3 of our more general result, to follow.

Now consider the more general case that the agents with non-empty access in M' are serial in a submodel M^{ser} of M that contains s, with domain S^{ser}. In other words: at least all agents who finally have consistent beliefs in some states, initially have consistent beliefs in all states. The construction above will no longer work: if the actual state is not considered possible by an agent, then that agent has empty access in actual state (s, s') of $(M \otimes \mathsf{U}'_r)$, but not in s' in M'. But if we relax the precondition $\delta_{(M,s)}$, for the point s of (M, s), to the disjunction $\bigvee_{u \in S^{\mathsf{ser}}} \delta_{(M,u)}$, that will now carry along the serial subdomain S^{ser}, the construction will work because an agent can then always imagine *some* state wherein the update has been executed, even it that is not the actual state. This indeed completes the construction.

Definition 3.1 (Update for arbitrary change)**.** Given finite epistemic models $M = (S, R, V)$ and $M' = (S', R', V')$ for the same sets of agents and atoms. Assume that all agents with non-empty access in M' are serial in M^{ser} containing s. The update for arbitrary change $(\mathsf{U}'', (s, s')) = ((\mathsf{E}'', \mathsf{R}'', \mathsf{pre}'', \mathsf{post}''), (s, s'))$ is defined as (for arbitrary agents a and arbitrary relevant atoms p):

$$
\begin{array}{lcl}
\mathsf{E}'' & = & S' \\
(t', u') \in \mathsf{R}''(a) & \text{iff} & (t', u') \in R'(a) \\
\mathsf{pre}''(t') & = & \bigvee_{u \in S^{\mathsf{ser}}} \delta_{(M,u)} \\
\mathsf{post}''(t')(p) & = & \begin{cases} \top & \text{if } t' \in V'(p) \\ \bot & \text{otherwise} \end{cases}
\end{array}
$$

The epistemic state $(M \otimes \mathsf{U}'', (s, s'))$ is bisimilar to (M', s'), which is the desired result. It will typically not be isomorphic: $M \otimes \mathsf{U}''$ can be seen as consisting of a number of copies of M' (namely $|S^{\mathsf{ser}}|$ copies) 'with the accessibility relations just right to establish the bisimulation'. One copy may not be enough, namely when the state t in M to which that copy corresponds, lacks access for some agents. This access will then also be 'missing' between the states of $(\{t\} \times S')$. But because of seriality one of

the other M' copies will now make up for this lack: there is a $u \in S^{\mathsf{ser}}$ such that $(t, u) \in R(a)$, which will establish access when required, as in the proof of the following proposition.

Proposition 3.2. Given (M, s), (M', s'), and U'' as in Definition 3.1. Then $\mathfrak{R} : ((M \otimes \mathsf{U}''), (s, s')) \leftrightarrows (M', s')$ by way of, for all $t \in S^{\mathsf{ser}}$ and $t' \in S'$: $\mathfrak{R}(t, t') = t'$.

Proof. Let $R^{\otimes}$ be the accessibility relation and $V^{\otimes}$ the valuation in $(M \otimes \mathsf{U}'')$.

atoms: For an arbitrary relevant atom p: $(t, t') \in V^{\otimes}(p)$ iff $(M, t) \models \mathsf{post}''(t')(p)$, and by definition of post'' we have that $(M, t) \models \mathsf{post}''(t')(p)$ iff $t' \in V'(p)$. Irrelevant atoms do not change value.

forth: Let $((t_1, t_1'), (t_2, t_2')) \in R^{\otimes}(a)$ and $((t_1, t_1'), t_1') \in \mathfrak{R}$. From $((t_1, t_1'), (t_2, t_2')) \in R^{\otimes}(a)$ follows $(t_1', t_2') \in R'(a)$. By definition of $\mathfrak{R}$ we also have $((t_2, t_2'), t_2') \in \mathfrak{R}$.

back: Let $((t_1, t_1'), t_1') \in \mathfrak{R}$ and $(t_1', t_2') \in R'(a)$. As M^{ser} is serial for a, and $t_1 \in S^{\mathsf{ser}}$, there must be a t_2 such that $(t_1, t_2) \in R(a)$. As $(M, t_2) \models \bigvee_{t \in \mathrm{dom}(M^{\mathsf{ser}})} \delta_{(M,t)}$ (because t_2 is one of those t) we have that $(t_2, t_2') \in \mathrm{dom}(M \otimes \mathsf{U}'')$. From that, $(t_1, t_2) \in R(a)$, and $(t_1', t_2') \in R'(a)$, follows that $((t_1, t_1'), (t_2, t_2')) \in R^{\otimes}(a)$. By definition of $\mathfrak{R}$ we also have $((t_2, t_2'), t_2') \in \mathfrak{R}$. Q.E.D.

Note that we keep the states outside the serial submodel M^{ser} out of the bisimulation. Without the seriality constraint the 'back' condition of the bisimilarity cannot be shown: given a $((t_1, t_1'), t_1') \in \mathfrak{R}$ and $(t_1', t_2') \in R'(a)$, but where t_1 has no outgoing arrow for a, the required a-accessible pair from (t_1, t_1') does not exist. A special case of Proposition 3.2 is the corollary already referred to during the initial two-step construction, that achieves even isomorphy:

Corollary 3.3. Given (M, s), (M', s'), and U'_r as above. Assume that M is a bisimulation contraction. Then $(M \otimes \mathsf{U}'_r) \cong M'$.

Proof. In this special case we have that $(t, t') \in \mathrm{dom}(M \otimes \mathsf{U}'_r)$ iff $(M, t) \models \mathsf{pre}'(t')$ iff $(M, t) \models \delta_{(M,s)}$ for the point s of (M, s). As the last is only the case when $t = s$ (as M is a bisimulation contraction), we end up with a domain consisting of all pairs (s, t') for all $t' \in S'$, a 1-1-correspondence. The bisimulation $\mathfrak{R}$ above becomes the isomorphism $\mathfrak{I}(s, t') = t'$. Q.E.D.

A different wording of Proposition 3.2 is that for arbitrary finite epistemic states (M, s) and (M', s') also satisfying the serial submodel constraint, there is an update (U, e) transforming the first into the second. A final appealing way to formulate this result is:

Corollary 3.4. Given are a finite epistemic state (M, s) and a satisfiable formula φ. If all agents occurring in φ have non-trivial beliefs in state s of M, then there is an update *realizing* φ, i.e., there is a (U, e) such that $(M, s) \models \langle \mathsf{U}, \mathsf{e} \rangle \varphi$.

Using completeness of the logic, this further implies that all consistent formulas can be realized in any given finite model. We find this result both weak and strong: it is strong because any conceivable (i.e., using the same propositional letters and set of agents) formal specification can be made true whatever the initial information state. At the same time, it is weak: the current information state does apparently not give any constraints on future developments of the system, or, in the opposite direction, any clue on the sequence of events resulting in it; the ability to change the value of atomic propositions arbitrarily gives too much freedom. Of course, if one restricts the events to *specific protocols* (such as legal game moves [15], and for a more general treatment see [11]), the amount of change is constrained.

AGM belief revision and belief update. Our results on arbitrary belief change seem related to the postulate of success in AGM belief revision [1]. AGM belief revision corresponds to epistemic change, and AGM (in their terminology) belief update [27] corresponds to ontic change (unfortunately, in the AGM community 'update' means something far more specific than what we mean by that term). Given this correspondence we can achieve only expansion by epistemic change, and not proper revision; and the combination of ontic and epistemic change can be seen as a way to make belief update result in belief revision. Apart from this obvious interpretation of epistemic and ontic change, one can also view our result that all consistent formulas can be realized, differently: a consequence of this is that for arbitrary consistent φ and ψ there is an update (U, e) such that $[a]\varphi \to \langle \mathsf{U}, \mathsf{e} \rangle [a]\psi$. In AGM terms: if φ is believed, then there is a way to revise that into belief of ψ, regardless of whether $\varphi \wedge \psi$ is consistent or not. In other words: revision with ψ is always successful. That suggests that our way of achieving that result by combining epistemic and ontic change might somehow simulate standard AGM belief revision. Unfortunately it is immediately clear that we allow far too much freedom for the other AGM postulates to be fulfilled. It is clearly not a *minimal* change, for example. So walking further down this road seems infeasible.

3.2 Postconditions true and false only

The postconditions for propositional atoms can be entirely simulated by the postconditions true or false for propositional atoms. For a simple example, the public assignment $p := \varphi$ can be simulated by a two-point update $\mathsf{e} \text{——} A \text{——} \mathsf{f}$ (i.e., a nondeterministic event where all agents in A cannot distinguish e from f) such that $\mathsf{pre}(\mathsf{e}) = \varphi$, $\mathsf{post}(\mathsf{e})(p) = \top$, $\mathsf{pre}(\mathsf{f}) = \neg\varphi$,

$\mathsf{post}(\mathsf{f})(p) = \bot$. In the public assignment $(p := \varphi, q := \psi)$ to two atoms p and q we would need a *four*-point update to simulate it, to distinguish all *four* ways to combine the values of two independent atoms.

The general construction consists of doing likewise in every event e of an update model. For each e we make as many copies as the cardinality of the powerset of the range of the postcondition associated with that event. Below, the set $\{0,1\}^{\mathrm{dom}(\mathsf{post}(\mathsf{e}))}$ represents that powerset.

Definition 3.5 (Update model $\mathsf{U}^{\top\!\!\perp}$)**.** Given is an update model $\mathsf{U} = (\mathsf{E}, \mathsf{R}, \mathsf{pre}, \mathsf{post})$. Then $\mathsf{U}^{\top\!\!\perp} = (\mathsf{E}^{\top\!\!\perp}, \mathsf{R}^{\top\!\!\perp}, \mathsf{pre}^{\top\!\!\perp}, \mathsf{post}^{\top\!\!\perp})$ is a *normal update model* with

- $\mathsf{E}^{\top\!\!\perp} = \bigcup_{\mathsf{e}\in\mathsf{E}}\{(\mathsf{e}, f) \mid f \in \{0,1\}^{\mathrm{dom}(\mathsf{post}(\mathsf{e}))}\}$
- $((\mathsf{e}, f), (\mathsf{e}', f')) \in \mathsf{R}^{\top\!\!\perp}(a)$ iff $(\mathsf{e}, \mathsf{e}') \in \mathsf{R}(a)$
- $\mathsf{pre}^{\top\!\!\perp}(\mathsf{e}, f) = \mathsf{pre}(\mathsf{e}) \wedge \bigwedge_{f(p)=1} \mathsf{post}(\mathsf{e})(p) \wedge \bigwedge_{f(p)=0} \neg\mathsf{post}(\mathsf{e})(p)$
- $\mathsf{post}^{\top\!\!\perp}(\mathsf{e}, f)(p) = \begin{cases} \top & \text{if } f(p) = 1 \\ \bot & \text{if } f(p) = 0 \end{cases}$

Proposition 3.6. Given an epistemic model $M = (S, R, V)$ and an update model $\mathsf{U} = (\mathsf{E}, \mathsf{R}, \mathsf{pre}, \mathsf{post})$ with normal update model $\mathsf{U}^{\top\!\!\perp}$ defined as above. Then $(M \otimes \mathsf{U}) \underline{\leftrightarrow} (M \otimes \mathsf{U}^{\top\!\!\perp})$.

Proof. We show that the relation $\mathfrak{R} : (M \otimes \mathsf{U}) \underline{\leftrightarrow} (M \otimes \mathsf{U}^{\top\!\!\perp})$ defined as

$$((s, \mathsf{e}), (s, \mathsf{e}, f)) \in \mathfrak{R} \text{ iff } (M, s) \models \mathsf{pre}^{\top\!\!\perp}(\mathsf{e}, f)$$

is a bisimulation. Below, the accessibility relations in $(M \otimes \mathsf{U})$ and $(M \otimes \mathsf{U}^{\top\!\!\perp})$ are also written as $R(a)$.

atoms

Let (s, e, f) be a state in the domain of $(M \otimes \mathsf{U}^{\top\!\!\perp})$. We have to show that for all atoms p, $(M, s) \models \mathsf{post}(\mathsf{e})(p) \leftrightarrow \mathsf{post}^{\top\!\!\perp}(\mathsf{e}, f)(p)$. From the definition of $\mathsf{post}^{\top\!\!\perp}$ it follows that

$$\mathsf{post}^{\top\!\!\perp}(\mathsf{e}, f)(p) \text{ iff } \ f(p) = 1 \ .$$

From $(M, s) \models \mathsf{pre}^{\top\!\!\perp}(\mathsf{e}, f)$ and the definition of $\mathsf{pre}^{\top\!\!\perp}$ follows that

$$(M, s) \models \mathsf{post}(\mathsf{e})(p) \text{ iff } \ f(p) = 1 \ .$$

Therefore

$$(M, s) \models \mathsf{post}(\mathsf{e})(p) \leftrightarrow \mathsf{post}^{\top\!\!\perp}(\mathsf{e}, f)(p) \ .$$

forth
Assume that $((s,\mathsf{e}),(s',\mathsf{e}')) \in R(a)$ and that $((s,\mathsf{e}),(s,\mathsf{e},f)) \in \mathfrak{R}$. Let $f' : \mathrm{dom}(\mathsf{post}(\mathsf{e}')) \to \{0,1\}$ be the function such that

$$f'(p) = \begin{cases} 1 & \text{if } (M,s') \models \mathsf{post}(\mathsf{e}')(p) \\ 0 & \text{otherwise} \end{cases}$$

Therefore $(M,s') \models \mathsf{pre}^{\top\!\!\perp}(\mathsf{e}',f')$. Therefore $((s',\mathsf{e}'),(s',\mathsf{e}',f')) \in \mathfrak{R}$. From $((s,\mathsf{e}),(s',\mathsf{e}')) \in R^{\top\!\!\perp}(a)$ follows $(s,s') \in R(a)$ and $(\mathsf{e},\mathsf{e}') \in \mathsf{R}(a)$. From $(\mathsf{e},\mathsf{e}') \in \mathsf{R}(a)$ and the definition of access $R^{\top\!\!\perp}$ follows $((\mathsf{e},f),(\mathsf{e}',f')) \in R^{\top\!\!\perp}(a)$. From $(s,s') \in R(a)$ and $((\mathsf{e},f),(\mathsf{e}',f')) \in R^{\top\!\!\perp}(a)$ follows $((s,\mathsf{e},f),(s',\mathsf{e}',f')) \in R(a)$.

back
Suppose $((s,\mathsf{e}),(s,\mathsf{e},f)) \in \mathfrak{R}$ and $(s,\mathsf{e},f),(s',\mathsf{e}',f') \in R(a)$. From the last follows $(s,s') \in R(a)$ and $((\mathsf{e},f),(\mathsf{e}',f')) \in \mathsf{R}^{\top\!\!\perp}(a)$, therefore also $(\mathsf{e},\mathsf{e}') \in \mathsf{R}(a)$. Therefore $((s,\mathsf{e}),(s,\mathsf{e}')) \in R(a)$. Just as in the case of **forth** it is established that $((s',\mathsf{e}'),(s',\mathsf{e}',f')) \in \mathfrak{R}$.

Q.E.D.

Corollary 3.7. The logic of change with postconditions true and false only is equally expressive as the logic of change with arbitrary postconditions.

Although it is therefore possible to use postconditions true and false only, this is highly unpractical in modelling actual situations: the descriptions of updates become cumbersome and lengthy, and lack intuitive appeal.

A transformation result similar to that in Proposition 3.6 can *not* be established for the logic with only singleton update models, i.e., the logic of public announcements and public assignments (as in [28]). If public assignments could only be to true and to false, then updates with assignments always result in models wherein the assigned atoms are true *throughout* the model, or false *throughout* the model. Therefore, there is no transformation of, e.g., $\underline{p}\!\!-\!\!-\!\!-\!\!\neg p$ into $p\!\!-\!\!-\!\!-\!\!\underline{\neg p}$ using public assignments and public announcements only. The construction above results in a *two-event* update model, that is not a singleton.

A transformation result as in Proposition 3.6 immediately gives an expressivity result as in Corollary 3.7 for the languages concerned. It is also tempting to see such a transformation result as a different kind of expressivity result. In two-sorted languages such as the one we consider in this paper one can then distinguish between the expressivity of two kinds of syntactic objects. A formula (φ) corresponds to a class of models that satisfy that formula, and a modality (α) corresponds to a relation on the class of models. The result is stated in terms of the expressivity of formulas, but

it is also a result about the expressivity of modalities. These two kinds of expressivity are not necessarily linked. One can have logics with the same expressivity of formulas, that have different expressivity of modalities and vice versa.

3.3 Single assignments only

Consider the update model e——a——f for a single agent a and for two atoms p_1 and p_2 such that in e, if φ_1 then $p_1 := \varphi_2$ and $p_2 := \varphi_3$, and in f, if φ_4 then $p_1 := \varphi_5$ and $p_2 := \varphi_6$. Can we also do the assignments one by one? In other words, does this update correspond to a sequence of updates consisting of events g in which at most one atom is assigned a value: the cardinality of $\mathrm{dom}(\mathsf{g})$ is at most 1. This is possible! *First* we 'store' the value (in a given model (M, s) wherein this update is executed) of all preconditions and postconditions in fresh atomic variables, by public assignments. This can be in arbitrary order, so we do it in the order of the φ_i. This is the sequence of six public assignments $q_1 := \varphi_1$, $q_2 := \varphi_2$, $q_3 := \varphi_3$, $q_4 := \varphi_4$, $q_5 := \varphi_5$, and $q_6 := \varphi_6$. Note that such public assignments do not change the structure of the underlying model. *Next* we execute the original update but without postconditions. This is e'——a——f' with $\mathsf{pre}(\mathsf{e}') = \mathsf{pre}(\mathsf{e}) = \varphi_1$ and $\mathsf{pre}(\mathsf{f}') = \mathsf{pre}(\mathsf{f}) = \varphi_4$ and with $\mathsf{post}(\mathsf{e}') = \mathsf{post}(\mathsf{f}') = \mathsf{id}$. Note that q_1 remains true whenever e' was executed, because q_1 was set to be true whenever φ_1 was true, the precondition of both e and e'. Similarly, q_4 remains true whenever f' was executed. We have now arrived at the final structure of the model, just not at the proper valuations of atoms.

Finally, the postconditions are set to their required value, *conditional* to the execution of the event with which they are associated. Agent a must not be aware of those conditions (the agent cannot distinguish between e' and f'). Therefore we cannot model this as a public action. The way out of our predicament is a number of two-event update models, namely one for each postcondition of each event in the original update. One of these two events has as its precondition the fresh atom associated with an event in the original update, and the other event its negation, and agent a cannot distinguish between both. The four required updates are

- e_1—a—e'_1 with in e_1, if q_1 then $p_2 := q_2$ and in e'_1, if $\neg q_1$ then id
- e_2—a—e'_2 with in e_2, if q_1 then $p_3 := q_3$ and in e'_2, if $\neg q_1$ then id
- e_3—a—e'_3 with in e_3, if q_4 then $p_5 := q_5$ and in e'_3, if $\neg q_4$ then id
- e_4—a—e'_4 with in e_4, if q_4 then $p_6 := q_6$ and in e'_4, if $\neg q_4$ then id

Now, we are done. These four final updates do not change the structure of the model, when executed. Therefore, now having set the postconditions right, the composition of all these constructs is *isomorphic* to the original

update model! The general construction is very much as in this simple example.

Definition 3.8 (Update model $\mathsf{U}^{\mathsf{one}}$)**.** Given an update model $\mathsf{U} = (\mathsf{E}, \mathsf{R}, \mathsf{pre}, \mathsf{post})$, update model $\mathsf{U}^{\mathsf{one}}$ is the composition of the following update models: *First* perform $\Sigma_{\mathsf{e}\in\mathsf{E}}|\mathrm{dom}(\mathsf{post}(\mathsf{e})) + 1|$ public assignments for fresh variables $q, \ldots,$ namely for each $\mathsf{e} \in \mathsf{E}$, $q^{\mathsf{e}}_0 := \mathsf{pre}(\mathsf{e})$, and for all $p_1, \ldots, p_n \in \mathrm{dom}(\mathsf{post}(\mathsf{e}))$, $q^{\mathsf{e}}_1 := \mathsf{post}(\mathsf{e})(p_1), \ldots, q^{\mathsf{e}}_n := \mathsf{post}(\mathsf{e})(p_n)$. *Then* execute U but with identity ('trivial') postconditions, i.e., execute $\mathsf{U}' = (\mathsf{E}, \mathsf{R}, \mathsf{pre}, \mathsf{post}')$ with $\mathsf{post}'(\mathsf{e}) = \mathsf{id}$ for all $\mathsf{e} \in \mathsf{E}$. *Finally*, execute $\Sigma_{\mathsf{e}\in\mathsf{E}}|\mathrm{dom}(\mathsf{post}(\mathsf{e}))|$ two-event update models with universal access for all agents wherein for each event just one of its postconditions is set to its required value, by way of the auxiliary atoms. For example, for $\mathsf{e} \in \mathsf{E}$ as above the first executed update is e_1——A——e_2 with in e_1, if q^{e}_0, then $p_1 := q^{\mathsf{e}}_1$, and in e_2, if $\neg q^{\mathsf{e}}_0$ then id.

The following proposition will be clear without proof:

Proposition 3.9. Given epistemic model M and update model U executable in M. Then $\mathsf{U}^{\mathsf{one}}$ is isomorphic to U, and $(M \otimes \mathsf{U}^{\mathsf{one}})$ is isomorphic to $(M \otimes \mathsf{U})$.

This result brings our logic closer to the proposals in [5, 16] wherein only one atom is simultaneously assigned a value. The relation to other proposals will be discussed in Section 5.

4 Card game actions

In this section we apply the logic to model multi-agent system dynamics in the general settings of various game actions for card players, such as showing, drawing, and swapping cards. The precise description of card game dynamics is a prerequisite to compute optimal strategies to play such games [13, 18]. Card deals are also frequently used as standard representation for cryptographic bit exchange protocols [22, 34], where communications/transmissions should from our current perspective be seen as public announcements and/or assignments.

Consider a deck of two Wheat, two Flax, and two Rye cards (w, x, y). Wheat, Flax and Rye are called the *commodities*. Three players Anne, Bill, and Cath (a, b, and c) each draw two cards from the stack. Initially, given a deal of cards, it is common knowledge what the deck of cards is, that all players hold two cards, and that all players (only) know their own cards. For the card deal where Anne holds a Wheat and a Flax card, Bill a Wheat and a Rye card, and Cath a Flax and a Rye card, we write $wx.wy.xy$, and so on. As the cards in one's hand are unordered, $wx.wy.xy$ is the same deal of cards as $xw.wy.xy$, but for improved readability we will always list cards in a hand in alphabetical order. There are certain game actions that result

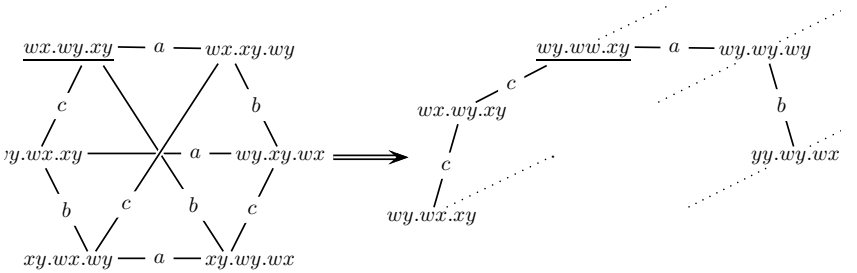


FIGURE 2. On the left is the game state after the cards have been dealt and Anne received Wheat and Flax, Bill received Wheat and Rye, and Cath received Flax and Rye. On the right is part of the game state that results if Anne trades her Wheat card for Bill's Rye card: only states resulting from trading Wheat for Wheat, and (what really happened) Wheat for Rye, are present. The actual deal of cards is underlined. In the figures, assume reflexivity and transitivity of access. The dotted lines suggest that some states are indistinguishable for Cath from yet other states but not present in the picture.

in players exchanging cards. This is called *trading* of the corresponding commodities. Players attempt to get two cards of the same suit. That is called establishing a *corner* in a commodity. Subject to game rules that are non-essential for our exposition, the first player to declare a corner in any commodity, wins the game. For example, given deal $wx.wy.xy$, after Anne swaps her Wheat card for Bill's Rye card, Bill achieves a corner in Wheat, and wins. Of course, players can already achieve a corner when the cards are dealt. This six-card scenario is a simplification of the 'Pit' card game that simulates the trading pit of a stock exchange [31, 13, 15]; the full game consists of 74 cards: 8 commodities of each 9 cards, and two special cards.

An initial game state wherein players only know their own cards and nobody has won the game yet, can be modelled as an epistemic state. There are six such card deals. Assume that the actual deal is $wx.wy.xy$. The (hexagonally shaped) epistemic state $(\mathsf{Pit}, wx.wy.xy)$ in Figure 2 pictures what players know about each other. All six card deals have to occur to describe the epistemic information present, e.g., Anne cannot distinguish actual deal $wx.wy.xy$ from deal $wx.xy.wy$, the other deal wherein Anne holds Wheat and Flax. But if the deal had been $wx.xy.wy$, Bill could not have distinguished *that* deal from $wy.xy.wx$, wherein Anne holds Wheat and Rye. Therefore, Anne considers it possible that Bill considers it possible that she holds Rye, even though this is actually not the case.

The event wherein Anne and Bill swap one of their cards involves both epistemic and ontic change. Assume that, given deal $wx.wy.xy$,

Anne swaps her Wheat card for Bill's Rye card.

This informal description is not specific enough to be modelled as an update.

In the first place, the role of Cath in the event is not specified. The event where Cath is unaware of the swap taking place, is different from the event where Cath observes the swap and where all agents are commonly aware of this. If Cath had been unaware of the event, she would be mistaken about the actual state of the world. For example, she would incorrectly still believe that neither Anne nor Bill has a corner in a commodity, whereas Bill holds two Wheat cards after the trade (and we assume that he has not yet so declared). It is hard to conceive of such a scenario as a *game*: even in imperfect information games, such as Pit, a basic condition of *fairness* must be fulfilled for players to be able to act rationally. This means that all events should at least be partially observable, and that 'what actually happened' should be considered a possibility for all players. We therefore assume, for now, that Cath learns that Anne and Bill swap one of their cards, but not which card it is. (The 'private swap' will be modelled later, as another example.)

Anne and Bill's roles in the event are also underspecified. Anne may knowingly *choose* a card to hand to Bill (the obvious interpretation), or blindly *draw* one of her cards to hand to Bill. The latter is not obvious, given this description, but becomes more so if we see it as Bill drawing (therefore blindly) one of Anne's cards. For now, assume the obvious. Another specification issue is that we may think of Bill as receiving Anne's Wheat card facedown and only then, in a subsequent action, picking it up. From our modelling perspective, Bill already can be said to *own* the card after he has been handed it, but before he has picked it up he does not yet *know* that he owns it. We first assume that players immediately 'see' the card they are being traded (in this case maybe not the most obvious choice, but the simplest one to model). In other words: Anne and Bill *jointly* learn the new ownership of both cards.

To describe this multi-agent system and its dynamics, assume a propositional language for three agents a, b, c and with atoms u_a^n expressing that Anne holds n cards of suit u. For example, w_a^2 expresses that Anne holds two Wheat cards. In the event where Anne and Bill swap Wheat for Rye, Anne gets one *less* Wheat card, Bill gets one *more* Wheat card, Bill gets one *less* Rye card, and Anne gets one *more* Rye card. In the update model we have to distinguish separate events for each card deal wherein this swap can take place, i.e., corresponding to $wx.wy.xy$, $wx.xy.wy$, and $wy.xy.wx$ (in general this depends on a feature of the local states of the card swapping agents only, namely for both agents on the number of Wheat and Rye cards in their hands, in this specific case that information is sufficient to determine the entire card deal). In case the card deal is $wx.wy.xy$ the precondition

and postcondition are

$$\text{If } (w_a^1 \wedge y_a^0 \wedge w_b^1 \wedge y_b^1), \text{ then}$$
$$w_a^1 := \bot \text{ and } w_a^0 := \top \text{ and } w_b^1 := \bot \text{ and } w_b^2 := \top$$
$$\text{and } y_a^0 := \bot \text{ and } y_a^1 := \top \text{ and } y_b^1 := \bot \text{ and } y_b^0 := \top.$$

We name this event $\text{swap}_{ab}^{wx.wy.xy}(w, y)$. If two cards of the same suit are swapped, a simpler description is sufficient. For example, the event wherein Anne and Bill swap Wheat given deal $wx.wy.xy$ is described as $\text{swap}_{ab}^{wx.wy.xy}(w, w)$ with (the same) precondition and (empty) postcondition

$$\text{If } (w_a^1 \wedge y_a^0 \wedge w_b^1 \wedge y_b^1), \text{ then } \varnothing.$$

From the point of view of an actual card deal, there are always four different ways to exchange a single card, namely for each agent either the one or the other card. All of these are clearly different for Anne and Bill, because they either give or receive a different card (we assumed that they know which card they give and see which card they receive). None of these are different for Cath. For different card deals, card swapping events are indistinguishable if those deals were indistinguishable. For example, the event where (Anne and Bill swap Wheat and Rye given $wx.wy.xy$) is indistinguishable for Anne from the event where (Anne and Bill swap Wheat and Rye given $wx.xy.wy$), because card deals $wx.wy.xy$ and $wx.xy.wy$ are the same for Anne.

Therefore, the update model for Anne swapping her Wheat card for Bill's Rye card consists of 24 events. The preconditions and postconditions of the events are as above. The accessibility relations are defined as, for deals $d, d' \in \text{dom}(\mathsf{Pit}) = \{wx.wy.xy, wx.xy.wy, \dots\}$ and cards $q, q', q_1, q_1' \in \{w, x, y\}$, and accessibility relations $R(a), R(b), R(c)$ in the epistemic model Pit:

$$(\text{swap}_{ab}^{d}(q, q'), \text{swap}_{ab}^{d'}(q_1, q_1')) \in \mathsf{R}(a) \text{ iff } (d, d') \in R(a), q = q_1 \text{ and } q' = q_1'$$
$$(\text{swap}_{ab}^{d}(q, q'), \text{swap}_{ab}^{d'}(q_1, q_1')) \in \mathsf{R}(b) \text{ iff } (d, d') \in R(b), q = q_1 \text{ and } q' = q_1'$$
$$(\text{swap}_{ab}^{d}(q, q'), \text{swap}_{ab}^{d'}(q_1, q_1')) \in \mathsf{R}(c) \text{ iff } (d, d') \in R(c)$$

We name the update model *Swap*. The event of Anne and Bill swapping Wheat for Rye has therefore been modelled as update $(\mathit{Swap}, \text{swap}_{ab}^{d}(w, y))$. The result of executing this update model in epistemic state $(\mathsf{Pit}, wx.wy.xy)$ has the same structure as the update model, as the preconditions are unique for a given state, and as access between events in the update model copies that in the epistemic state. It has been partially visualized in Figure 2. An intuitive way to see the update and the resulting structure in Figure 2, is as a restricted product of the Pit model and *nine* card swapping

events $\mathrm{swap}(q, q_1)$, namely for each combination of the three different cards. Figure 2 then shows just two of those copies, namely for $\mathrm{swap}(w, y)$ and $\mathrm{swap}(w, w)$. For example, the event $\mathrm{swap}(w, y)$ 'stands for' the three events $\mathrm{swap}_{ab}^{wx.wy.xy}(w, y)$, $\mathrm{swap}_{ab}^{wy.xy.wx}(w, y)$, and $\mathrm{swap}_{ab}^{wx.xy.wy}(w, y)$.

Why did we not define such $\mathrm{swap}(q, q_1)$ as updates in their own right, in the first place? Although intuitive, this is not supported by our modelling language. We would like to say that the postconditions are 'Anne gets *one less* Wheat card, and Bill gets *one more* Wheat card,' and similarly for Rye. But instead, we only *can* demand that in case Bill already had a Wheat card (extra precondition), *then* he now has two, etc. Incidentally, we can also add non-deterministic choice to the update language by notational abbreviation, as $[\alpha \cup \beta]\varphi \leftrightarrow ([\alpha]\varphi \wedge [\beta]\varphi)$ (this corresponds to taking the *union* of the epistemic state transformations induced by α and β). We can then define, in the update language, $\mathrm{swap}(w, y) = \mathrm{swap}_{ab}^{wx.wy.xy}(w, y) \cup \mathrm{swap}_{ab}^{wy.xy.wx}(w, y) \cup \mathrm{swap}_{ab}^{wx.xy.wy}(w, y)$.

The case where Anne does not choose her card but Bill blindly draws one of Anne's can also be modelled as an update. The accessibility for Anne then expresses that she is unaware which of her cards has been drawn:

$$(\mathrm{swap}_{ab}^{d}(q, q'), \mathrm{swap}_{ab}^{d'}(q_1, q_1')) \in \mathsf{R}(a) \quad \text{iff} \quad (d, d') \in R(a) \text{ and } q' = q_1'$$

This is somewhat counterintuitive when we still suppose that Anne observes which card she receives from Bill. (We'd have to imagine Bill blindly drawing one of Anne's cards, Anne putting her remaining card facedown on the table, and receiving the card Bill gives her faceup.) A more realistic setting is then that Bill draws one of Anne's card and 'pushes the card he gives to Anne facedown towards her'. At that stage Anne can already be said to own the card, but not yet to know that. All four swapping actions for a given deal are indistinguishable for Anne (as they were and still are for Cath).

Yet another event is where Anne chooses a card to show to Bill, and receives Bill's card facedown (before she picks it up). Access is now

$$(\mathrm{swap}_{ab}^{d}(q, q'), \mathrm{swap}_{ab}^{d'}(q_1, q_1')) \in \mathsf{R}(a) \quad \text{iff} \quad (d, d') \in R(a) \text{ and } q = q_1$$

Obviously, all these variables can be applied to Bill as well.

Picking up a card. The action of picking up a card after it has been handed to you, has another description (see, using another dialect of the language, [15]). One aspect is interesting to observe in the current context. Imagine that given deal $wx.wy.xy$ after Anne and Bill swapping Wheat and Rye, Bill receives the card facedown in the following way: after having laid down his remaining card (a Wheat card) facedown on the table, Anne puts her Wheat card facedown on top of it or under it in a way that Bill cannot

see that. He now picks up his two cards. How does Bill know which of the two Wheat cards that he then holds, is the received card? He does not know, but he does not care either. By looking at his cards after the swap, he effectively *learns* that he holds two Wheat cards (which was already true after having received the card), and after that event he then *knows* that he holds two Wheat cards. A neat way to express that he learns the suit of the card he received, is to say that there is a suit for which he learns to have one more card than he knows. This makes sense in the general Pit game setting wherein one only is allowed to trade a certain number of cards of the same suit. This is formalized in our setting by an update (U, e) with an event with precondition $\mathsf{pre}(\mathsf{e}) = w_b^2 \wedge \neg[b]w_b^2$ (and empty postcondition), as a result of which Bill then knows to hold two Wheat cards, i.e., $[\mathsf{U}, \mathsf{e}][b]w_b^2$.

Private swap. The event where Anne and Bill swap Wheat for Rye but where Cath is unaware of the event is modelled by a four-event update with events $\mathrm{swap}_{ab}^{wx.wy.xy}(w, y)$, $\mathrm{swap}_{ab}^{wy.wx.xy}(w, y)$, $\mathrm{swap}_{ab}^{wx.wy.xy}(w, y)$ and skip, such that for Anne and Bill access among the swap events is as already discussed (including all variations), but where 'Cath thinks nothing happens': for the deals d in the three swap events: $(\mathrm{swap}_{ab}^{d}(w, y), \mathsf{skip}) \in R(c)$, and $(\mathsf{skip}, \mathsf{skip}) \in R(a), R(b), R(c)$.

5 Comparison to other approaches

Action model logic. Dynamic modal operators for ontic change, in addition to similar operators for epistemic change, have been suggested in various recent publications. As far as we know it was first mentioned by Baltag, Moss, and Solecki as a possible extension to their *action model logic* (for epistemic change), in [5]. This was by example only and without a language or a logic. A precise quotation of all these authors say on ontic change may be in order:

> *Our second extension concerns the move from actions as we have been working them to actions which change the truth values of atomic sentences. If we make this move, then the axiom of Atomic Permanence*[5] *is no longer sound. However, it is easy to formulate the relevant axioms. For example, if we have an action α which effects the change $p := p \wedge \neg q$, then we would take an axiom $[\alpha]p \leftrightarrow (PRE(\alpha) \rightarrow p \wedge \neg q)$. Having made these changes, all the rest of the work we have done goes through. In this way, we get a completeness theorem for this logic.* [5, p. 24]

The logic that we present here is a realization of their proposal, and we can confidently confirm that the authors were correct in observing that "all the rest (...) goes through". To obtain such theoretical results the

[5] I.e., $[\alpha]p \leftrightarrow (PRE(\alpha) \rightarrow p)$, [5, p. 15].

notion of *simultaneous* postconditions (assignments) for a finite subset of atomic propositional letters is essential; this feature is not present in [5] (but introduced in [20]).

In a later proposal by Baltag [2] a fact changing action *flip*P is proposed that changes ('flips') the truth value of an atom P, with accompanying axioms (for the proper correspondent action α resembling a single-pointed action model) $[\alpha]p \leftrightarrow (\mathsf{pre}(\alpha) \rightarrow \neg p)$ if "p changes value (flips) in α", and otherwise $[\alpha]p \leftrightarrow (\mathsf{pre}(\alpha) \rightarrow p)$ [2, p. 29]. The approach is restricted to ontic change where the truth value of atoms flips. In the concluding section of [2], the author defers the relation of this proposal to a general logic of ontic and epistemic change to the future.

Recent work in dynamic epistemics. More recently, in a Multi-Agent Systems application-driven line of research [16, 15] assignments are added to the relational action language of [12] but without providing an axiomatization. In this setting only change of *knowledge* is modelled and not change of belief, i.e., such actions describe transformation of S5 models only, such as Kripke models corresponding to interpreted systems.

A line of research culminating in *Logics of communication and change* [10] also combines epistemic and ontic change. It provides a more expressive setting for logical dynamics than our approach. The logic presented here is a sublogic of LCC. In [10] the focus is on obtaining completeness via so-called reduction axioms for dynamic epistemic logics, by extending the basic modal system to PDL. Our treatment of postconditions, also called substitutions, stems from [20]. In the current paper we focus on specific semantic results, and, as said, we use a dynamic epistemic 'dialect', not full PDL.

A recent contribution on combining *public* ontic and epistemic change, including detailed expressivity results for different combinations of static and dynamic modalities, is found in [28]. Our work uses a similar approach to ontic events but describes more complex than public events: the full generality of arbitrarily complex events involves exchange of cards among subgroups of the public, and other events with a 'private' (as opposed to public) character.

Finally, a general dynamic modal logic is presented in [33], where ontic changes are also studied. The semantics of this logic uses tree-like structures, and fixed points are introduced in the language to be able to reason about updates.

Belief revision. An independent recent line of investigation combining epistemic with ontic change arises from the belief revision community. Modelling *belief revision*, i.e., epistemic change, by dynamic operators is an old idea going back to Van Benthem [7]. In fact, this is one of the two original publications—together with [32]—that starts the area of dynamic epistemic

logic. For an overview of such matters see [3, 14, 17]. But modelling ontic change—known as *belief update* [27]—in a similar, dynamic modal, way, including its interaction with epistemic change, is only a recent focus of ongoing research by Herzig and collaborators and other researchers based at the *Institut de Recherche en Informatique de Toulouse* (IRIT) [25, 26, 29]. Their work sees the execution of an event as so-called *progression* of information, and reasoning from a final information state to a sequence of events realizing it as *regression*—the last obviously relates to planning. The focus of progression and regression is the change of the *theory* describing the information state, i.e., the set of all true, or believed, formulas. As already mentioned in Section 3, the results for arbitrary belief change in Proposition 3.2 and following corollaries can potentially be applied to model belief update in the AGM tradition.

Interpreted systems. In a way, dynamic epistemic logics that combine epistemic and ontic change reinvent results already obtained in the interpreted systems community by way of knowledge-based programs [30, 21]: in that setting, ontic and epistemic change are integrated. Let us point out some correspondences and differences, using the setting of van der Meyden's [30]. This work investigates the implementation of knowledge-based programs. The transition induced by an update between epistemic states, in our approach, corresponds exactly to a step in a run in an interpreted system that is the implementation of a knowledge-based program; the relation between both is explicit in van der Meyden's notion of the *progression structure*. Now the dynamic epistemic approach is both more general and more restrictive than the interpreted systems approach. It is more restrictive because dynamic epistemics assumes perfect recall and synchronicity. This assumption is implicit: it is merely a consequence of taking a state transition induced by an update as primitive. But the dynamic epistemic approach is also somewhat more general: it does not assume that accessibility relations for agents are equivalence relations, as in interpreted systems. In other words, it can also be used to model other epistemic notions than knowledge, such as introspective belief and even weaker notions.

Knowledge-based programs consist of joint actions $\langle a_e, a_1, \ldots, a_n \rangle$ where a_e is an action of the environment and where $a_1, \ldots, a_n$ are *simultaneous* actions by the agents 1 to n. An agent a acts according to conditions of the form '*if* φ' *do* a', *if* φ'' *do* a'', ...' etc. Let us overlook the aspect that conditions φ' have the form of known formulas (by agent a). Still, such statements *look* familiar to our alternative format for what goes on in an event, as in '*for event* e*: if* φ, *then* $p_1 := \psi_1$, ..., *and* $p_n := \psi_n$.' (see after Definition 2.4 on page 91). This correspondence is not really there, but the similar format is still revealing. The different cases in a knowledge-based program are like the different events in an update model, and they

equally express non-determinism. This is a correspondence. There are also differences. In dynamic epistemics, the condition φ in '*if* φ*, then* $p_1 := \psi_1$*, ...*' is both an executability precondition *and* an observation. Inasmuch as it is an observation, it determines agent knowledge. In the interpreted systems approach, observations are (with of course reason) modelled as different from preconditions. The assignments such as $p_1 := \psi_1$ in the 'then' part of event descriptions are merely the ontic part of that event, with the 'if' part describing the epistemic part. Epistemic and ontic features *together* correspond to actions such as a' in '*if* φ' *do* a''. In the interpreted systems approach, epistemic and ontic features of information change are therefore not separately modelled, as in our approach.

6 Further research

An unresolved issue is whether updates can be described as compositions of purely epistemic events (preconditions only) and purely ontic events (postconditions only). In [25] it is shown for public events, for a different logical (more PDL-like) setting. Such a result would be in the line of our other normalization results, and facilitate comparison to related approaches. The result in Section 3.1 seems to suggest this is possible, since a transition from one epistemic model to another is achieved by an epistemic event followed by an ontic event. However, the method described in Section 3.1 is geared towards the original epistemic model. We would like a decomposition that is based solely on the update, which would work regardless of the particular epistemic model to which it is applied.

The logic can be applied to describe cryptographic bit exchange protocols, including protocols where keys change hands or are being sent between agents. The logic is very suitable for the description of protocols for computationally unlimited agents, such as described in the cited [22, 34]. Using dynamic logics may be an advantage given the availability of model checking tools for such logics, as e.g., the very versatile epistemic model checker DEMO [19] by van Eijck. The current version of DEMO only allows epistemic events. But van Eijck and collaborators are in the process of extending DEMO with assignments (postconditions), needed to model events that are both epistemic and ontic.

Our more fine-grained analysis of events may contribute to the description and verification of more complex protocols that also include non-public events. An example that demonstrates the applicability of the logic to the analysis of such protocols is the (solution of the) 'one hundred prisoners and a light bulb' riddle (see e.g., [35]), of which we have a detailed analysis in preparation that we consider of independent interest.

The results for 'arbitrary belief change' suggest yet another possibly promising direction. Under certain conditions arbitrary formulas are real-

izable. What formulas are still realizable if one restricts the events to those considered suitable for specific problem areas, such as forms of multi-agent planning? And given a desirable formula (a 'postcondition' in another sense of the word), what are the initial conditions such that a sequence of events realizes it? This is the relation to AI problems concerning regression as pointed out in the introductory section [25], and also to reasoning given specific protocols, such as always has been the emphasis for knowledge-based programs in the interpreted systems community [21], and as recently investigated in [11] in a dynamic epistemic context.

Acknowledgments

We thank the anonymous reviewers of this contribution to the volume. We also thank Johan van Benthem and Wiebe van der Hoek for their comments and their encouragement.

References

[1] C.E. Alchourrón, P. Gärdenfors & D. Makinson. On the logic of theory change: partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

[2] A. Baltag. A logic for suspicious players: Epistemic actions and belief updates in games. *Bulletin of Economic Research*, 54(1):1–45, 2002.

[3] A. Baltag, H.P. van Ditmarsch & L.S. Moss. Epistemic logic and information update. In J.F.A.K. van Benthem & P. Adriaans, eds., *Handbook on the Philosophy of Information*. Elsevier, Amsterdam. Forthcoming.

[4] A. Baltag & L.S. Moss. Logics for epistemic programs. *Synthese*, 139(2):165–224, 2004.

[5] A. Baltag, L.S. Moss & S. Solecki. The logic of public announcements, common knowledge, and private suspicions. Tech. rep., Centrum voor Wiskunde en Informatica, Amsterdam, 1999. CWI Report SEN-R9922.

[6] J. Barwise & L.S. Moss. *Vicious Circles: On the Mathematics of Non-Wellfounded Phenomena*. CSLI Publications, Stanford, 1996.

[7] J.F.A.K. van Benthem. Semantic parallels in natural language and computation. In H.-D. Ebbinghaus, J. Fernandez-Prida, M. Garrido, D. Lascar & M.R. Artalejo, eds., *Logic Colloquium '87*. North-Holland, Amsterdam, 1989.

[8] J.F.A.K. van Benthem. Dynamic odds and ends. Tech. rep., University of Amsterdam, 1998. *ILLC Publication* ML-1998-08.

[9] J.F.A.K. van Benthem. 'One is a lonely number': on the logic of communication. In Z. Chatzidakis, P. Koepke & W. Pohlers, eds., *Logic Colloquium '02*, vol. 27 of *Lecture Notes in Logic*. Association for Symbolic Logic, 2006.

[10] J.F.A.K. van Benthem, J. van Eijck & B.P. Kooi. Logics of communication and change. *Information and Computation*, 204(11):1620–1662, 2006.

[11] J.F.A.K. van Benthem, J.D. Gerbrandy & E. Pacuit. Merging frameworks for interaction: DEL and ETL. In D. Samet, ed., *Proceedings of the 11th Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pp. 72–81. UCL Presses Universitaires de Louvain, 2007.

[12] H.P. van Ditmarsch. Descriptions of game actions. *Journal of Logic, Language and Information*, 11(3):349–365, 2002.

[13] H.P. van Ditmarsch. Some game theory of Pit. In C. Zhang, H.W. Guesgen & W.-K. Yeap, eds., *PRICAI 2004: Trends in Artificial Intelligence, 8th Pacific Rim International Conference on Artificial Intelligence, Auckland, New Zealand, August 9–13, 2004, Proceedings*, vol. 3157 of *Lecture Notes in Computer Science*, pp. 946–947. Springer, 2004.

[14] H.P. van Ditmarsch. Belief change and dynamic logic. In J. Delgrande, J. Lang, H. Rott & J.-M. Tallon, eds., *Belief Change in Rational Agents: Perspectives from Artificial Intelligence, Philosophy, and Economics*, no. 05321 in Dagstuhl Seminar Proceedings. IBFI, Schloss Dagstuhl, 2005.

[15] H.P. van Ditmarsch. The logic of Pit. *Synthese*, 149(2):343–375, 2006.

[16] H.P. van Ditmarsch, W. van der Hoek & B.P. Kooi. Dynamic epistemic logic with assignment. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 05)*, pp. 141–148. ACM Inc., New York, 2005.

[17] H.P. van Ditmarsch, W. van der Hoek & B.P. Kooi. *Dynamic Epistemic Logic*, vol. 337 of *Synthese Library*. Springer, 2007.

[18] S. Druiven. *Knowledge Development in Games of Imperfect Information*. Master's thesis, University of Groningen, 2002.

[19] J. van Eijck. Dynamic epistemic modelling. Tech. rep., Centrum voor Wiskunde en Informatica, Amsterdam, 2004. CWI Report SEN-E0424.

[20] J. van Eijck. Guarded actions. Tech. rep., Centrum voor Wiskunde en Informatica, Amsterdam, 2004. CWI Report SEN-E0425.

[21] R. Fagin, J. Halpern, Y. Moses & M. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, MA, 1995.

[22] M. Fischer & R. Wright. Bounds on secret key exchange using a random deal of cards. *Journal of Cryptology*, 9(2):71–99, 1996.

[23] J. Gerbrandy & W. Groeneveld. Reasoning about information change. *Journal of Logic, Language, and Information*, 6(2):147–169, 1997.

[24] D. Harel, D. Kozen & J. Tiuryn. *Dynamic Logic*. MIT Press, Cambridge, MA, 2000. Foundations of Computing Series.

[25] A. Herzig & T. De Lima. Epistemic actions and ontic actions: A unified logical framework. In J.S. Sichman, H. Coelho & S.O. Rezende, eds., *Advances in Artificial Intelligence — IBERAMIA-SBIA 2006, 2nd International Joint Conference, 10th Ibero-American Conference on AI, 18th Brazilian AI Symposium, Ribeirão Preto, Brazil, October 23–27, 2006, Proceedings*, vol. 4140 of *Lecture Notes in Computer Science*, pp. 409–418. Springer, 2006.

[26] A. Herzig, J. Lang & P. Marquis. Action progression and revision in multiagent belief structures, 2005. Manuscript presented at Sixth Workshop on Nonmonotonic Reasoning, Action, and Change (NRAC 2005).

[27] H. Katsuno & A. Mendelzon. On the difference between updating a knowledge base and revising it. In J.F. Allen, R. Fikes & E. Sandewall, eds., *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning (KR'91). Cambridge, MA, USA, April 22–25, 1991*, pp. 387–394. Morgan Kaufmann, 1991.

[28] B.P. Kooi. Expressivity and completeness for public update logics via reduction axioms. *Journal of Applied Non-Classical Logics*, 17(2):231–253, 2007.

[29] N. Laverny. *Révision, mises à jour et planification en logique doxastique graduelle*. Ph.D. thesis, Institut de Recherche en Informatique de Toulouse (IRIT), Toulouse, France, 2006.

[30] R. van der Meyden. Constructing finite state implementations of knowledge-based programs with perfect recall. In L. Cavedon, A. Rao & W. Wobcke, eds., *Intelligent Agent Systems, Theoretical and Practical Issues (based on PRICAI'96)*, vol. 1209 of *Lecture Notes in Computer Science*, pp. 135–151. Springer, 1997.

[31] Pit game rules. See `http://www.hasbro.com/common/instruct/pit.pdf`.

[32] J. Plaza. Logics of public communications. In M. Emrich, M. Pfeifer, M. Hadzikadic & Z. Ras, eds., *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems: Poster Session Program*, pp. 201–216. Oak Ridge National Laboratory, 1989. ORNL/DSRD-24.

[33] G.R. Renardel de Lavalette. Changing modalities. *Journal of Logic and Computation*, 14(2):253–278, 2004.

[34] A. Stiglic. Computations with a deck of cards. *Theoretical Computer Science*, 259(1–2):671–678, 2001.

[35] W. Wu. 100 prisoners and a lightbulb, 2001. Manuscript.

Social Laws and Anti-Social Behaviour

Wiebe van der Hoek
Mark Roberts
Michael Wooldridge

Department of Computer Science
University of Liverpool
Liverpool L69 7ZF, United Kingdom
{wiebe,mark,mjw}@csc.liv.ac.uk

Abstract

We extend our work that uses ATL to reason about social laws. In a system with social laws, every agent is supposed to refrain from performing certain forbidden actions. Rather than assuming that all agents abide to the law, we reason about what happens if certain agents do, i.e., they act socially, while others don't. In particular, we are interested in the strategic abilities, under such mixed conditions. We also compare our approach with one in which labels are added to ATL that record whether agents behaved socially or not.

1 Introduction

A multiagent system [15] on the one hand aims at treating the agents as autonomous entities, whose behaviour should not be over-specified or too constrained, while at the same time one wants this system to achieve, or maintain certain objectives. As a consequence, one of the defining problems in multiagent systems research is that of *coordination*—managing the interdependencies between the actions of multiple interacting agents [1, 15]. There are broadly two techniques to approach this. Online techniques aim to equip agents with the ability to dynamically coordinate their activities, for example by explicitly reasoning about coordination at run-time. In contrast, offline techniques aim at developing a coordination regime at design-time, and hardwiring this regime into a system for use at run-time. There are arguments in favour of both approaches: the former is potentially more flexible, and may be more robust against unanticipated events, while the latter approach benefits from offline reasoning about coordination, thereby reducing the run-time decision-making burden on agents [15].

One prominent approach to "offline" coordination is the *social laws* paradigm, introduced largely through the work of Shoham, Tennenholtz, and Moses [12, 11, 13, 14]. A social law can be understood as a set of rules

Giacomo Bonanno, Wiebe van der Hoek, Michael Wooldridge (eds.). *Logic and the Foundations of Game and Decision Theory (LOFT 7).* Texts in Logic and Games 3, Amsterdam

imposed upon a multiagent system with the goal of ensuring that some desirable global behaviour will result. Social laws work by *constraining* the behaviour of the agents in the system—by *forbidding* agents from performing certain actions in certain circumstances. Limiting the agents' abilities as this may sound, this may also open new opportunities for them: assuming that other agents abide to the norms, may imply that an individual can in fact achieve more. Shoham and others investigated a number of issues surrounding the development and use of social laws, including the computational complexity of their synthesis, and the possibility of the development of social laws or conventions by the agents within the system themselves.

In [5], we extended this social laws framework. In particular, we argued that *Alternating-time Temporal Logic* (ATL) provides a rich and natural technical framework within which to investigate social laws and their properties. In this framework, a social law consists of two parts: an objective of the law, written as an ATL specification, and a behavioural constraint, which is a function that for every action returns the set of states where that action is forbidden from being performed. The objective of the law represents what the society aims to achieve by adopting the law, and the behavioural constraint corresponds to the requirements that the law places on the members of society. In [4] we then added an epistemic flavour to this ATL-based approach to social laws, enabling one not only to express that an agent should behave in certain ways *given his information*, but also that certain information should emerge by following a social law.

In this paper we extend our social laws framework further. In our previous work [5, 4], it is assumed that once social laws are imposed on the system, all the agents will abide by these laws. However, this does not seem to be the most realistic way of modelling social laws. Certainly in human societies, just because laws are imposed does not mean that all members of the society will follow these laws. In this paper we do away with the need for this assumption and give agents the choice of whether to follow these laws or not. We make a distinction between agents acting *physically* and agents acting *socially*. Acting physically—we use this in lack of a better term we know—corresponds to performing any action that is physically possible to be performed, in the sense of being possible by the system description. Acting socially then corresponds to performing any action from a subset of these physical actions, known as the social actions. Social actions are those that are consistent with the social laws imposed on the system. An example scenario is in the case of traffic laws. Traffic lights are used to coordinate the flow of traffic on our roads to ensure no collisions occur when cars cross each other's paths. Cars are only *allowed* to move from the traffic lights when the light is green. Acting socially (and physically) would correspond to only moving when the lights are green. However, acting phys-

ically (but not socially) could correspond to moving when the lights are red. This would be an illegal action, but still a physically possible action. It is important to note that all social actions are also physically possible actions.

Agents now have both physical and social strategies, where the set of social strategies is always a subset of the physical strategies. So after giving the agents these extra possibilities, we need to extend the logic to be able to reason about whether an agent or coalition of agents is able to act socially or physically (of course if an agent is able to act physically, this agent can choose to only perform social strategies if it desires). We introduce two logical languages: in *Social* ATL (SATL), one can express properties like the following: "Even if the coalition abides by the laws, and all the other agents neglect them, our coalition can achieve a particular temporal property", and in *Social* ATEL (SATEL) one can also refer to informational attitudes of the agents, like in: "the server knows that, where all the clients are acting socially, then eventually each granted write-permission will terminate". Finally, we investigate an alternative approach for expressing properties of systems that refer to whether the agents are acting socially or physically. Rather than using the logical language, Social ATL, we see to what extent we can capture the same notions using only ATL and ATL*.

This paper is structured as follows: We first introduce the semantic structures that our model is based on. We call these Social Action-based Alternating (Epistemic) Transition Systems (SAAETS and SAATS). We then introduce our logical languages, Social ATEL and Social ATL, and give its semantics. In Section 4 we introduce a case study known as the Alternating Bit Protocol and go on to investigate various interesting properties of this model. In Section 5 we try to capture similar properties in an alternative framework and find direct equivalences between the two. Finally, we conclude in Section 6.

2 Semantic Structures

In this section we introduce the semantic structures our model is based upon. The structures we use are called Social Action-Based Alternating Epistemic Transition Systems (SAAETSs). Our structures are most similar to the AAETSs we introduced in [4]. However, our structures differ from those in several ways. Instead of having one action pre-condition function, ρ, we now have both a physical action precondition function, ρ, and a legal action precondition function, ℓ. Also, where the emphasis in [5] is on implementing a social law on a system, we do not consider such updates: rather, we assume the ℓ function is given, constraining the set of possible transitions, and we are not concerned about which social *goal* or *objective* such constraints are supposed to achieve. Rather, this framework is used to investigate social laws at the system level. Hence the system designer is able to see which

properties hold depending on who follows the social laws.

Formally, an SAAETS is a $(2n + 8)$-tuple

$$\langle Q, q_0, \mathsf{Ag}, \mathsf{Ac}_1, \ldots, \mathsf{Ac}_n, \sim_1, \ldots, \sim_n, \rho, \ell, \tau, \Phi, \pi \rangle \quad \text{where}$$

- Q is a finite, non-empty set of *states*;
- $q_0 \in Q$ is the designated *initial state* of the system;
- $\mathsf{Ag} = \{1, \ldots, n\}$ is a finite, non-empty set of *agents*;
- Ac_i is a finite, non-empty set of *actions*, for each $i \in \mathsf{Ag}$, where $\mathsf{Ac}_i \cap \mathsf{Ac}_j = \varnothing$ for all $i \neq j \in \mathsf{Ag}$. With $\mathsf{Ac}_{\mathsf{Ag}}$ we mean the set of all actions: $\mathsf{Ac}_{\mathsf{Ag}} = \bigcup_{i \in \mathsf{Ag}} \mathsf{Ac}_i$. Moreover, J_{Ag} is the set of *joint actions* $j = \langle a_1, \ldots, a_n \rangle$, with action $j_i = a_i$ in Ac_i $(i \leq n)$;
- $\sim_i \subseteq Q \times Q$ is an epistemic accessibility relation for each agent $i \in \mathsf{Ag}$. Each $\sim_i$ must be an equivalence relation.
- $\rho : \mathsf{Ac}_{\mathsf{Ag}} \to 2^Q$ is a *physical action precondition function*, which for each action $\alpha \in \mathsf{Ac}_{\mathsf{Ag}}$ defines the set of states $\rho(\alpha)$ from which α may be *physically* performed; and,
- $\ell : \mathsf{Ac}_{\mathsf{Ag}} \to 2^Q$ is a *legal action precondition function*, which for each action $\alpha \in \mathsf{Ac}_{\mathsf{Ag}}$ defines the set of states $\ell(\alpha)$ from which α may be *physically* and *legally* performed. We require that for all $\alpha \in \mathsf{Ac}_{\mathsf{Ag}}$, $\ell(\alpha) \subseteq \rho(\alpha)$.
- $\tau : Q \times J_{\mathsf{Ag}} \to Q$ is a partial *system transition function*, which defines the state $\tau(q, j)$ that would result by the performance of j from state q—note that, as this function is partial, not all joint actions are possible in all states (cf. the physical action pre-condition function above).
- Φ is a finite, non-empty set of *atomic propositions*; and
- $\pi : Q \to 2^\Phi$ is an interpretation function, which gives the set of primitive propositions satisfied in each state: if $p \in \pi(q)$, then this means that the propositional variable p is satisfied (equivalently, true) in state q.

As with AATSs, SAAETSs must satisfy two coherence constraints: Firstly, *non-triviality [11]*: Agents always have at least one *legal* action: $\forall q \in Q, \forall i \in \mathsf{Ag}, \exists \alpha \in \mathsf{Ac}_i$ s.t. $q \in \ell(\alpha)$. Secondly, *consistency*: The ρ and τ functions agree on actions that may be performed: $\forall q, \forall j \in J_{\mathsf{Ag}}, (q, j) \in \operatorname{dom} \tau$ iff $\forall i \in \mathsf{Ag}, q \in \rho(j_i)$. Sometimes we are not interested in knowledge and leave out the accessibility relations: such a system is called a SAATS.

Example 2.1 (The Train Example)**.** We refer to this system as Train. There are two trains, one of which (E) is Eastbound, the other of which (W) is Westbound, each occupy their own circular track. At one point, both tracks pass through a narrow tunnel—a crash will occur if both trains are in the tunnel at the same time.

We model each train $i \in \mathsf{Ag} = \{E, W\}$ as an automaton that can be in one of three states: "away_i" (the initial state of the train); "$\mathsf{waiting}_i$" (waiting to enter the tunnel); and "in_i" (the train is in the tunnel). Initially, both trains are away. Each train $i \in \{E, W\}$ has two actions available. They can either move or idle. Idling causes no change in the train's state. If a train i moves while it is away_i, then it goes to a $\mathsf{waiting}_i$ state; moving while $\mathsf{waiting}_i$ causes a transition to an in_i state; and finally, moving while in_i causes a transition to away_i as long as the other train was not in the tunnel, while if both trains are in the tunnel, then they have crashed, and will idle indefinitely. In order to prevent this from happening, train E is not allowed to move when either W is waiting or in the tunnel, while ℓ forbids W to move when E is in. See Figure 1, where $\sim_i(q)$ is shorthand for $\{q' \mid q \sim_i q'\}$ ($i \in \{E, W\}$). So $\sim_i(q)$ are those states that look similar to q, for agent i. In our example, each train knows about itself where it is: away, waiting or in the tunnel.

Given an agent $i \in \mathsf{Ag}$ and a state $q \in Q$, we denote the *physical options* available to i in q by $\rho\text{-}options(i, q) = \{\alpha \mid \alpha \in \mathsf{Ac}_i \text{ and } q \in \rho(\alpha)\}$ and, similarly, the *legal options* available to i in q by $\ell\text{-}options(i, q) = \{\alpha \mid \alpha \in \mathsf{Ac}_i \text{ and } q \in \ell(\alpha)\}$. Let Q^* be the set of finite sequences of elements of Q, with typical element $\mu, \nu \in Q^*$, and where $\mu \cdot q$ denotes concatenation of an element μ from Q^* with a state $q \in Q$. $Q^+ \subseteq Q^*$ collects all finite sequences of length at least 1. Now we can define a *physical strategy* and a *legal strategy* for an agent. A *physical strategy* for an agent $i \in \mathsf{Ag}$ is a function: $\gamma_i : Q^+ \to \mathsf{Ac}_i$ which must satisfy the constraint that $\gamma_i(\mu \cdot q) \in \rho\text{-}options(i, q)$ for all $\mu \in Q^*$ and $q \in Q$. A *legal strategy* for an agent $i \in \mathsf{Ag}$ is a function: $\delta_i : Q^+ \to \mathsf{Ac}_i$ which must satisfy the *legality* constraint that $\delta_i(\mu \cdot q) \in \ell\text{-}options(i, q)$ for all $\mu \in Q^*$ and $q \in Q$.

A *physical strategy profile* for a coalition $G = \{1, \ldots, k\} \subseteq \mathsf{Ag}$ is a tuple of physical strategies $\langle \gamma_1, \ldots, \gamma_k \rangle$, one for each agent $i \in G$. Similarly, a *legal strategy profile* for a coalition $G = \{1, \ldots, k\} \subseteq \mathsf{Ag}$ is a tuple of legal strategies $\langle \delta_1, \ldots, \delta_k \rangle$, one for each agent $i \in G$. By Γ_G, we denote the set of all *physical* strategy profiles for coalition G, whereas $\Delta_G \subseteq \Gamma_G$ denotes the set of all *legal* strategy profiles for G. By σ_G, we denote a strategy profile for coalition G where we are not concerned about whether the strategy profile is legal or not; if $\sigma_G \in \Gamma_{\mathsf{Ag}}$ and $i \in \mathsf{Ag}$, then we denote i's component of σ_G by σ_G^i. Given a coalition G, a grand coalition strategy profile σ_{Ag} can be considered as a tuple $\langle \sigma_G, \sigma'_{\bar{G}} \rangle$, where σ_G represents the choices made by

States and Initial States:

$Q = \{q_0, q_1, q_2, q_3, q_4, q_5, q_6, q_7, q_8\}$ Initial state q_0

Epistemic relations:

$\sim_E(q_0) = \{q_0, q_1, q_2\}$ $\sim_E(q_3) = \{q_3, q_5, q_6\}$ $\sim_E(q_4) = \{q_4, q_7, q_8\}$

$\sim_W(q_0) = \{q_0, q_3, q_4\}$ $\sim_W(q_1) = \{q_1, q_5, q_7\}$ $\sim_W(q_2) = \{q_2, q_6, q_8\}$

Agents, Actions, and Joint Actions:

$\mathsf{Ag} = \{E, W\}$ $\mathsf{Ac}_E = \{\mathsf{idle}_E, \mathsf{move}_E\}$ $\mathsf{Ac}_W = \{\mathsf{idle}_W, \mathsf{move}_W\}$

$J_{\mathsf{Ag}} = \{\underbrace{\langle \mathsf{idle}_E, \mathsf{idle}_W\rangle}_{j_0}, \underbrace{\langle \mathsf{idle}_E, \mathsf{move}_W\rangle}_{j_1}, \underbrace{\langle \mathsf{move}_E, \mathsf{idle}_W\rangle}_{j_2}, \underbrace{\langle \mathsf{move}_E, \mathsf{move}_W\rangle}_{j_3}\}$

Transitions/Physical action pre-conditions:

$$\tau(q_0, j) = \begin{cases} q_0 & \text{if } j = j_0 \\ q_1 & \text{if } j = j_1 \\ q_3 & \text{if } j = j_2 \\ q_5 & \text{if } j = j_3 \end{cases} \qquad \tau(q_4, j) = \begin{cases} q_4 & \text{if } j = j_0 \\ q_7 & \text{if } j = j_1 \\ q_0 & \text{if } j = j_2 \\ q_1 & \text{if } j = j_3 \end{cases}$$

$$\tau(q_1, j) = \begin{cases} q_1 & \text{if } j = j_0 \\ q_2 & \text{if } j = j_1 \\ q_5 & \text{if } j = j_2 \\ q_6 & \text{if } j = j_3 \end{cases} \qquad \tau(q_5, j) = \begin{cases} q_5 & \text{if } j = j_0 \\ q_6 & \text{if } j = j_1 \\ q_7 & \text{if } j = j_2 \\ q_8 & \text{if } j = j_3 \end{cases}$$

$$\tau(q_2, j) = \begin{cases} q_2 & \text{if } j = j_0 \\ q_0 & \text{if } j = j_1 \\ q_6 & \text{if } j = j_2 \\ q_3 & \text{if } j = j_3 \end{cases} \qquad \tau(q_6, j) = \begin{cases} q_6 & \text{if } j = j_0 \\ q_3 & \text{if } j = j_1 \\ q_8 & \text{if } j = j_2 \\ q_4 & \text{if } j = j_3 \end{cases}$$

$$\tau(q_3, j) = \begin{cases} q_3 & \text{if } j = j_0 \\ q_5 & \text{if } j = j_1 \\ q_4 & \text{if } j = j_2 \\ q_7 & \text{if } j = j_3 \end{cases} \qquad \tau(q_7, j) = \begin{cases} q_7 & \text{if } j = j_0 \\ q_8 & \text{if } j = j_1 \\ q_1 & \text{if } j = j_2 \\ q_2 & \text{if } j = j_3 \end{cases}$$

$$\tau(q_8, j_k) = q_8 \ (k = 0 \ldots 3)$$

Legal action precondition function:

$\ell(\mathsf{idle}_E) = Q$ $\ell(\mathsf{move}_E) = Q \setminus \{q_5, q_6\}$

$\ell(\mathsf{idle}_W) = Q$ $\ell(\mathsf{move}_W) = Q \setminus \{q_7\}$

Propositional Variables:

$\Phi = \{\mathsf{away}_E, \mathsf{away}_W, \mathsf{waiting}_E, \mathsf{waiting}_W, \mathsf{in}_E, \mathsf{in}_W\}$

Interpretation Function:

$\pi(q_0) = \{\mathsf{away}_E, \mathsf{away}_W\}$

$\pi(q_1) = \{\mathsf{away}_E, \mathsf{waiting}_W\}$

$\pi(q_2) = \{\mathsf{away}_E, \mathsf{in}_W\}$

$\pi(q_3) = \{\mathsf{waiting}_E, \mathsf{away}_W\}$

$\pi(q_4) = \{\mathsf{in}_E, \mathsf{away}_W\}$

$\pi(q_5) = \{\mathsf{waiting}_E, \mathsf{waiting}_W\}$

$\pi(q_6) = \{\mathsf{waiting}_E, \mathsf{in}_W\}$

$\pi(q_7) = \{\mathsf{in}_E, \mathsf{waiting}_W\}$

$\pi(q_8) = \{\mathsf{in}_E, \mathsf{in}_W\}$

FIGURE 1. The SAAETS for the trains scenario.

agents in G and $\sigma'_{\bar{G}}$ represents the choices made by all other agents in the system.

Given a strategy profile σ_{Ag} and non-empty sequence $\mu \cdot q \in Q^+$, let $\mathrm{out}(\sigma_{\mathsf{Ag}}, \mu \cdot q)$ denote the next state that will result by the members of Ag acting as defined by their components of σ_{Ag} for one step from $\mu \cdot q$. Formally, $\mathrm{out}(\sigma_{\mathsf{Ag}}, \mu \cdot q) = \tau(q, j) = q'$, where $\sigma^i_{\mathsf{Ag}}(\mu \cdot q) = j_i$ for $i \in \mathsf{Ag}$. Given a strategy profile σ_{Ag} for the grand coalition Ag, and a state $q \in Q$, we define $\mathsf{comp}(\sigma_{\mathsf{Ag}}, q)$ to be the run that will occur if every agent $i \in \mathsf{Ag}$ follows the corresponding strategy σ_i, starting when the system is in state $q \in Q$. Formally, $\mathsf{comp}(\sigma_{\mathsf{Ag}}, q) = \lambda = \lambda[0]\lambda[1]\ldots$ where $\lambda[0] = q$ and $\forall u \in \mathbb{N} : \lambda[u+1] = \mathrm{out}(\sigma_{\mathsf{Ag}}, \lambda[0]\lambda[1]\ldots\lambda[u])$.

Given an SAAETS, S, and the initial state of the system, q_0, we define a set of *socially reachable states* as follows:

$$\mathsf{sreach}(S) = \{q \mid \exists \delta_{\mathsf{Ag}} \in \Delta_{\mathsf{Ag}} \text{ and } \exists u \in \mathbb{N} \text{ s.t. } q = \mathsf{comp}(\delta_{\mathsf{Ag}}, q_0)[u]\}$$

The socially reachable states are all the states which are reachable when every agent in the system performs only social strategies.

3 Social ATEL

In this section we define the logical language used to express social laws in our framework. Our language is essentially an extension of ATEL [7]. In SATEL, we reason about physical and about social strategies. Our modalities are of the form $\langle\!\langle G \rangle\!\rangle^y_x$, where x and y denote which kind of strategies G and $\mathsf{Ag} \setminus G$, respectively, are allowed to use: only social strategies (denoted by s), or all their available ones (denoted by p). For example, the formula $\langle\!\langle G \rangle\!\rangle^s_p \Diamond \mathsf{goal}$ expresses that there exists a strategy for the coalition G, such that, no matter what the other agents do as long as they only follow social strategies, at some point in the future some goal state will occur. We can also require all the agents to act socially, e.g., $\langle\!\langle G \rangle\!\rangle^s_s \Box \neg \mathsf{fail}$, which expresses that G has a social strategy, such that, no matter what the other agents do, providing they act socially, the system will never enter a fail state. Finally, consider the nested formula, $\langle\!\langle G \rangle\!\rangle^s_s \bigcirc \langle\!\langle G \rangle\!\rangle^p_s \Box \varphi$, which reads: "$G$ can ensure, by using a social strategy, and assuming that all the others also act socially, that in the next state, G can ensure, again by acting socially, that even if the others from now on act non-socially, φ will always hold". Or, a bit more informally: "if we require G to act socially, and the others socially for at least the first step, but unconstrained thereafter, then G can guarantee that always φ".

ATEL adds knowledge operators K_i on top of ATL. However, on top of that we add operators to express more enhanced informational attitudes. First of all, now that the possibility of violating social laws exists, we define a notion of *social belief*, i.e., belief under the assumption that *all* agents

in the system are acting as they should do according to the social laws imposed. We introduce a belief operator SB_i, where $SB_i\varphi$ expresses that, if i assumes that everybody acts socially, he believes φ to be the case. For example, if we use φ to denote the fact that a car is stopped at a red traffic light, $SB_i\varphi$ means that agent i believes that a car is stopped at a red traffic light, under the assumption that all agents in the system are acting socially. So in all indistinguishable *social* states to agent i, a car is stopped at a red traffic light. An obvious difference with the standard notion of knowledge is that social belief is not veridical: it is perfectly well possible that agent i socially believes φ ($SB_i\varphi$) while $\neg\varphi$ at the same time.

Formally, the set of formulae, formed with respect to a set of agents Ag, and a set of primitive propositions Φ, is given by the following grammar:

$$\varphi \ ::= \ p \mid \neg\varphi \mid \varphi \vee \varphi \mid K_i\varphi \mid SB_i\varphi \mid \langle\langle G\rangle\rangle^y_x \bigcirc \varphi \mid \langle\langle G\rangle\rangle^y_x \Box\varphi \mid \langle\langle G\rangle\rangle^y_x \varphi \mathcal{U} \varphi$$

where $p \in \Phi$ is a propositional variable, $G \subseteq \mathsf{Ag}$ is a set of agents, $i \in \mathsf{Ag}$ is an agent, and x and y can be either p or s.

We now give the truth definition of Social ATEL formulae on an SAAETS S and a state q. For $G \subseteq \mathsf{Ag}$ and $x \in \{s, p\}$, let $\mathsf{Strat}(G, x) = \Gamma_G$ if $x = p$, and $\mathsf{Strat}(G, x) = \Delta_G$ if $x = s$.

$S, q \models p$	iff $p \in \pi(q)$ (where $p \in \Phi$)
$S, q \models \neg\varphi$	iff $S, q \not\models \varphi$;
$S, q \models \varphi \vee \psi$	iff $S, q \models \varphi$ or $S, q \models \psi$;
$S, q \models \langle\langle G\rangle\rangle^y_x \bigcirc \varphi$	iff $\exists \sigma_G \in \mathsf{Strat}(G, x)$ s.t. $\forall \sigma_{\bar{G}} \in \mathsf{Strat}(\bar{G}, y)$, if $\lambda = \mathsf{comp}(\langle \sigma_G, \sigma_{\bar{G}} \rangle, q)$, we have $S, \lambda[1] \models \varphi$;
$S, q \models \langle\langle G\rangle\rangle^y_x \Box\varphi$	iff $\exists \sigma_G \in \mathsf{Strat}(G, x)$ s.t. $\forall \sigma_{\bar{G}} \in \mathsf{Strat}(\bar{G}, y)$, if $\lambda = \mathsf{comp}(\langle \sigma_G, \sigma_{\bar{G}} \rangle, q)$, we have for all$n \in \mathbb{N}$, $S, \lambda[n] \models \varphi$;
$S, q \models \langle\langle G\rangle\rangle^y_x \varphi \mathcal{U} \psi$	iff $\exists \sigma_G \in \mathsf{Strat}(G, x)$ s.t. $\forall \sigma_{\bar{G}} \in \mathsf{Strat}(\bar{G}, y)$, if $\lambda = \mathsf{comp}(\langle \sigma_G, \sigma_{\bar{G}} \rangle, q)$, there is $k \in \mathbb{N}$ s.t. $S, \lambda[k] \models \psi$, and for all n with $0 \leq n < k$, we have $S, \lambda[n] \models \varphi$;
$S, q \models K_i\varphi$	iff for all q' such that $q \sim_i q' : S, q' \models \varphi$;
$S, q \models SB_i\varphi$	iff for all $q' \in \mathsf{sreach}(S)$ such that $q \sim_i q'$, we have $S, q' \models \varphi$.

Other connectives ("$\wedge$", "$\rightarrow$", "$\leftrightarrow$") are assumed to be defined as abbreviations in terms of $\neg, \vee$. $\langle\langle G\rangle\rangle^y_x \Diamond\varphi$ is shorthand for $\langle\langle G\rangle\rangle^y_x \top \mathcal{U} \varphi$. We write $\langle\langle i\rangle\rangle$ rather than $\langle\langle\{i\}\rangle\rangle$. Validity of φ, written $\models \varphi$ is defined as usual.

An *objective* formula is a purely propositional formula, with no knowledge, coalitional or temporal operators. The sublanguage obtained from SATEL by leaving out the epistemic and social belief operators is SATL. The language of ATL can be seen as being subsumed by SATL. Coalitional modalities in ATL are implicitly indexed with a p: the only type of strategy in ATL is physical. In ATL, one writes $\langle\!\langle G\rangle\!\rangle \bigcirc \varphi$ (corresponding to our $\langle\!\langle G\rangle\!\rangle^s_s \bigcirc \varphi$), and similarly for the other temporal operators. Note that in SATEL and ATL every temporal operator is immediately proceeded by a coalition modality. If one drops this constraint on ATL, the thus obtained language is called ATL*. Examples of ATL* formulas are $\langle\!\langle G\rangle\!\rangle(\,\Box(\varphi \rightarrow \diamondsuit\psi)$ ('G can cooperate such that whenever φ occurs, eventually ψ will happen') and $\langle\!\langle G\rangle\!\rangle(\bigcirc\varphi \vee \bigcirc\bigcirc\varphi \vee \bigcirc\bigcirc\bigcirc\varphi)$ ('G has a strategy that ensures that within three steps, φ').

In ATL, the cooperation modality $\langle\!\langle\rangle\!\rangle T$ (where T is a temporal formula, i.e., a formula of which the main operator is either $\bigcirc$, $\Box$ or $\mathcal{U}$) denotes a special case: it means that the empty set of agents has a strategy, such that, no matter what the other (i.e., *all*) agents do, T holds. In other words, no matter what the agents in Ag do, T. This resembles the CTL operator $\mathsf{A}T$: on every future path, T. Similarly, $\langle\!\langle \mathsf{Ag}\rangle\!\rangle T$ means that the grand coalition has a strategy such that, no matter what the empty coalition does, T. In CTL terminology: $\mathsf{E}T$, or, for some path, T. For SATEL this gives us the following. Since in $\langle\!\langle\rangle\!\rangle^y_x$, the constraint to play an x-type of strategy is a constraint for nobody, it does not really matter whether x is s or p. Similarly, in $\langle\!\langle \mathsf{Ag}\rangle\!\rangle^y_x T$ the constraint to play a y-type of strategy is a void restriction for $\mathsf{Ag} \setminus \mathsf{Ag}$, i.e., the empty set, so it does not matter whether x equals s or equals p. Summarising, we have

$$\langle\!\langle\rangle\!\rangle^y_s T \equiv \langle\!\langle\rangle\!\rangle^y_p T \text{ and } \langle\!\langle \mathsf{Ag}\rangle\!\rangle^s_x T \equiv \langle\!\langle \mathsf{Ag}\rangle\!\rangle^p_x T$$

As a convention, when having an empty coalition, we will only write $\langle\!\langle\rangle\!\rangle^s_s$ and $\langle\!\langle\rangle\!\rangle^p_p$, which is no restriction, given the equivalence above. Similarly, for the full coalition, we will only write $\langle\!\langle \mathsf{Ag}\rangle\!\rangle^s_s$ and $\langle\!\langle \mathsf{Ag}\rangle\!\rangle^p_p$.

Multiple $\langle\!\langle G\rangle\!\rangle^y_x \bigcirc$ operators are used as an abbreviation in the following way:

$$(\langle\!\langle G\rangle\!\rangle^y_x \bigcirc)^n \varphi \triangleq \begin{cases} \langle\!\langle G\rangle\!\rangle^y_x \bigcirc \varphi \text{ if } n = 1 \\ (\langle\!\langle G\rangle\!\rangle^y_x \bigcirc)(\langle\!\langle G\rangle\!\rangle^y_x \bigcirc)^{n-1}\varphi \text{ otherwise} \end{cases}$$

Social strategies Δ_G are a subset of the physical strategies Γ_G, and hence:

$$\models \langle\!\langle G\rangle\!\rangle^y_s T \rightarrow \langle\!\langle G\rangle\!\rangle^y_p T \text{ and } \models \langle\!\langle G\rangle\!\rangle^p_x T \rightarrow \langle\!\langle G\rangle\!\rangle^s_x T \tag{3.1}$$

where T here is an arbitrary temporal formula and x and y are variables over p and s. These properties express the following. The first,

$\langle\langle G\rangle\rangle^y_s T \rightarrow \langle\langle G\rangle\rangle^y_p T$ says that if a coalition G are able to enforce a property φ by adopting social strategies, then they can also enforce this same property when adopting physical strategies ('if you can enforce it nicely, you can enforce it anyhow'); and $\langle\langle G\rangle\rangle^p_x T \rightarrow \langle\langle G\rangle\rangle^s_x T$ can be interpreted as saying that if a coalition G are able to enforce a property φ when playing against an adversary who is able to use physical strategies, then they can also enforce this property when playing against the same adversary when this adversary is constrained to use only *social* strategies ('if you can beat an opponent when he can cheat, you can beat him when he plays by the rules').

4 Case Study

We now present a case study in order to demonstrate Social ATEL. The case study is known as "The Alternating Bit Protocol" and is adapted from [6]. In this scenario there are three agents, a sender S and a receiver R, who wish to communicate through a communication environment, represented as an agent E. The sender S owns a tape $X = \langle x_0, x_1, \ldots\rangle$, where each element x_i comes from a given alphabet, say $\mathsf{Alph} = \{0, 1\}$. The goal of the protocol is that the contents of this tape are sent to R, who writes what he receives on a tape Y that, after k elements of X have been received, will be $Y = \langle y_0, y_1 \ldots, y_k\rangle$. The sender only sends one item x_i at the time. The environment E determines whether messages are delivered or get lost: E does not alter the content of any message. And, although the environment is unreliable, it satisfies the fairness property that it will not loose messages forever: if an agent repeatedly sends the same message, eventually it will be delivered. We wish to design a protocol that satisfies the safety requirement that the receiver never prints incorrect bits, and the liveness requirement that every bit will eventually be written by the receiver onto Y.

The obvious solution to this problem is to use acknowledgements to let the sender know an element x_i has been received. So the sender would repeatedly send the current element x_i until eventually it receives an acknowledgement back from the receiver, at which point it would go on to send the next element x_{i+1}. The problem arises when the value of an item x_{i+1} is the same as x_i. The receiver does not know whether its acknowledgement for x_i has been received, so the receiver does not know whether the message he currently receives (with the value of x_{i+1}) is a repetition of x_i, or whether this bit is supposed to represent x_{i+1} on the tape. To overcome this problem, the sender can put more information on which element he sends by adding a "colour" to it: a 0 to every x_i for even i, and a 1 for every odd i. So the alphabet will now be $\mathsf{Alph}' = \{x.0 \mid x \in \mathsf{Alph}\} \cup \{x.1 \mid x \in \mathsf{Alph}\}$. We also need two acknowledgements: $\mathsf{ack0}$ and $\mathsf{ack1}$. Now the protocol works as follows: S first sends $x_0.0$. When it receives $\mathsf{ack0}$, it goes on to the next

state on the tape and sends $x_1.1$. R can see that this is a new message (since it has a different colour), and sends ack1 to acknowledge the receipt of it. S can also see that this is a different acknowledgement and starts to send $x_2.0$, and so on.

We model the scenario with an SAAETS called AltBit. We introduce variables with the following interpretation: $SSX = 1$ means that the last bit S sent was of type $x.1$; $SRA = 1$ indicates the last message received by S was ack1; $RRX = 1$ means the last bit R received was of type $x.1$; and $RSA = 1$ indicates the last message R sent was ack1. In the object language, atom ssx will mean that $SSX = 1$, and $\neg ssx$ denotes $SSX = 0$. Similarly for the other variables and corresponding atoms.

A state in our system is defined to be a tuple

$$qi = \langle SSX, SRA \mid RRX, RSA \rangle$$

where

- $SSX, SRA, RRX, RSA \in \{0, 1\} = \mathbb{B}$; and
- qi, where $1 \leq i \leq 16$, is the name of the state. This is just a decimal representation of the binary number the state corresponds to (e.g., $q3 = \langle 0, 0 \mid 1, 1 \rangle$). The initial state of the system is $q15 = \langle 1, 1 \mid 1, 1 \rangle$: no information about the first bit x_0 is assumed to be sent or received.

Given the nature of our states and the correspondence between variables and atomic propositions, we have: $ssx \in \pi(\langle SSX, SRA \mid RRX, RSA \rangle)$ iff $SSX = 1$, and similarly for the other variables. We are not interested in knowledge of E, but for S and R we assume that they know the contents of their own variables: if $q = \langle SSX, SRA \mid RRX, RSA \rangle$ and $q' = \langle SSX', SRA' \mid RRX', RSA' \rangle$ are two states, then $q \sim_S q'$ iff $SSX = SSX'$ & $SRA = SRA'$, and similarly for $\sim_R$.

In every state of the system, the sender has two physical actions available to it: send.0 and send.1, corresponding to sending a bit with colour 0 and sending a bit with colour 1, respectively. The receiver also has two physical actions available to it in every state of the system: sendack.0 and sendack.1, corresponding to sending an acknowledgement of a bit with colour 0 and colour 1, respectively. Actions of the environment are pairs (e_S, e_R), where e_S is either d_S or g_S (i.e., either deny or grant the request of the sender), and $e_R \in \{d_R, g_R\}$.

Concerning the transition function $\tau : Q \times J_{\mathsf{Ag}} \to Q$, we have the following:

$$\tau(\langle SSX, SRA \mid RRX, RSA \rangle, \langle a_S, a_R, a_E \rangle) = \langle SSX', SRA' \mid RRX', RSA' \rangle$$

where

$$SSX' = \begin{cases} 1 & \text{if } a_S = \mathsf{send}.1 \\ 0 & \text{if } a_S = \mathsf{send}.0 \end{cases}$$

$$SRA' = \begin{cases} SRA & \text{if } a_E = (\cdot, d_R) \\ 1 & \text{if } a_R = \mathsf{sendack}.1 \ \&\ a_E = (\cdot, g_R) \\ 0 & \text{if } a_S = \mathsf{sendack}.0 \ \&\ a_E = (\cdot, g_R) \end{cases}$$

$$RRX' = \begin{cases} RRX & \text{if } a_E = (d_S, \cdot) \\ 1 & \text{if } a_S = \mathsf{send}.1 \ \&\ a_E = (g_S, \cdot) \\ 0 & \text{if } a_S = \mathsf{send}.0 \ \&\ a_E = (g_S, \cdot) \end{cases}$$

$$RSA' = \begin{cases} 1 & \text{if } a_R = \mathsf{sendack}.1 \\ 0 & \text{if } a_R = \mathsf{sendack}.0 \end{cases}$$

This transition function reflects the fact that both S and R are free to decide which message they send: S can choose the value of SSX' and R chooses that of RSA'. However, whether a message arrives depends on E as well. For instance, in order for a new state to have $SRA' = 1$, saying that the last acknowledgment that S received was an $\mathsf{ack}1$, either this value was just sent by R and this was granted by E, or S had received $SRA' = 1$ in a previous state, and in the meantime he did not receive any update.

Regarding the legal transitions, of course, the idea is that S should alternate $x.0$'s, once known to be received, with $x.1$'s, and R is expected not to acknowledge receipt of a bit he did not receive. Let q be $\langle SSX, SRA \mid RRX, RSA \rangle$. Then:

$$\begin{array}{lcllcl} q \in \ell(\mathsf{send}.1) & \Leftrightarrow & SRA = 0 & q \in \ell(\mathsf{sendack}.1) & \Leftrightarrow & RRX = 1 \\ q \in \ell(\mathsf{send}.0) & \Leftrightarrow & SRA = 1 & q \in \ell(\mathsf{sendack}.0) & \Leftrightarrow & RRX = 0 \end{array}$$

For E, rather than putting a legality constraint, we require a fairness constraint on its behaviour. We say that a computation $\lambda = q_0 q_1 \cdots$ is fair if

- there are infinitely many indices n for which there is an action a_S for the sender and an action a_R for the receiver and some $e_R \in \{g_R, d_R\}$ such that $q_{n+1} = \tau(q_n, \langle a_S, a_R, (g_S, e_R) \rangle)$.
- there are infinitely many indices n for which there is an action a_S for the sender and an action a_R for the receiver and some $e_S \in \{g_S, d_S\}$ such that $q_{n+1} = \tau(q_n, \langle a_S, a_R, (e_S, g_R) \rangle)$.

In words: a fair computation is one in which both the request of the sender and that of the receiver are granted infinitely often. Although fairness

is a property of computations, we here can connect it to the legality of strategies of E. The legal strategies Δ_E of E are all strategies that guarantee that the resulting computation is fair, no matter what R and S do.

A *socially necessary fact* φ is a fact which should be true no matter what, providing all the agents in the system act in a social manner. A SNF φ is defined as follows:

$$\mathrm{SNF}(\varphi) \equiv \langle\langle\rangle\rangle^s_s \Box\varphi$$

4.1 Properties of the model

Let us discuss some properties, and verify whether they hold, either in our system AltBit, or in AltBit, $q15$. Recall that $\langle\langle\rangle\rangle^p_p \Box\varphi$ means that φ is true no matter what the agents do, while $\langle\langle\rangle\rangle^s_s \Box\psi$ says that ψ is true, as long as everybody acts socially. For instance, under what circumstances is it the case that if S sends an item (say $x_i.1$), R will eventually receive it? For this, 'only' E has to act socially: we have

$$\langle\langle\rangle\rangle^p_p \Box(ssx \rightarrow \langle\langle E\rangle\rangle^p_s \Diamond rrx) \tag{4.1}$$

This is good, and indeed, without demanding that E acts socially, he could indeed spoil it: we have $\langle\langle\rangle\rangle^p_p \Box((ssx \wedge \neg rrx) \rightarrow \langle\langle E\rangle\rangle^s_p \Box\neg rrx)$, saying, that when ssx is true, E can guarantee, even if the others act socially, that this message will never arrive. Returning to (4.1), it is not hard to see that E's next action is partially determined if he wants to ensure $\Diamond rrx$: if E does not immediately grant the request of S (i.e., E plays (d_S, d_R) or (d_S, g_R)), and we allow S to act physically, it can from any next state on decide to send $x.0$ (where, socially speaking, it should send $x.1$ again), and then, no matter which strategy E chooses, rrx may never be true again. In other words, we do *not* have

$$\langle\langle\rangle\rangle^p_p \Box(ssx \rightarrow \langle\langle\rangle\rangle^s_s \bigcirc(rrx \vee \langle\langle E\rangle\rangle^p_s \Diamond rrx)) \tag{4.2}$$

i.e., it is not the case that under all social strategies, in the next state the message has either been received, or else the environment can still guarantee it will be received eventually. (The counterexample is the earlier given strategies of E and S.)

If E and S both are social, we have the following positive result, which says that if $x.1$ is sent, then E and S can socially guarantee that this remains true until it is received, and moreover, it will be received under any behaviour, as long as E and S act socially:

$$\langle\langle\rangle\rangle^p_p \Box(ssx \rightarrow (\langle\langle E, S\rangle\rangle^p_s (ssx \,\mathcal{U}\, rrx) \wedge \neg\langle\langle E, S\rangle\rangle^p_s \Box\neg rrx)) \tag{4.3}$$

Now suppose R has received a message with odd parity: rrx. How can we derive that S will know about this? In other words, for which group G

and attitudes x and y do we have $\mathsf{AltBit}, q_{15} \models \langle\langle\rangle\rangle^p_p \Box(rrx \rightarrow \langle\langle G\rangle\rangle^y_x \Diamond sra)$? In case $x = y = p$ it holds iff $G \supseteq \{R, E\}$: the receiver and environment are needed to ensure that a received bit by the receiver is acknowledged to the sender. This only says that G *can* do it, however, and we also have $\langle\langle\rangle\rangle^p_p \Box(rrx \rightarrow \langle\langle G\rangle\rangle^p_p \Diamond \neg sra)$, even for $G = \{R\}$ and for $G = \{E\}$: in both cases, G can ensure that S does not get acknowledged properly. For that, illegal actions are needed, however, and indeed we have $\langle\langle\rangle\rangle^p_p \Box(rrx \rightarrow \neg\langle\langle R, E\rangle\rangle^p_s \Box \neg sra)$: If E and R act legally, it is impossible for them to not have a message from the sender ever acknowledged. Indeed, we even have $\langle\langle\rangle\rangle^s_s(rrx\,\mathcal{U}\,(rrx \wedge rsa))$: if everybody acts socially, received messages will be acknowledged and not overwritten in the meantime.

Let us write $\mathrm{SLT}(\varphi, \psi)$, ('$\varphi$ socially leads to ψ') if the following holds: $\varphi \rightarrow \langle\langle\rangle\rangle^s_s(\varphi\,\mathcal{U}\,\psi)$. That is, if φ is true, then, if everybody acts socially, it will stay true until some point where ψ is true. We then have:

$$
\begin{aligned}
\langle\langle\rangle\rangle^s_s \Box \quad (\quad & \mathrm{SLT}((\neg ssx \wedge \neg srx), (ssx \wedge \neg srx)) \ \wedge && (4.4)\\
& \mathrm{SLT}((ssx \wedge \neg srx), (rrx \wedge rsx)) \ \wedge && (4.5)\\
& \mathrm{SLT}((rrx \wedge rsx), (ssx \wedge srx)) \ \wedge && (4.6)\\
& \mathrm{SLT}((ssx \wedge srx), (\neg ssx \wedge srx)) \ \wedge && (4.7)\\
& \mathrm{SLT}((\neg ssx \wedge srx), (\neg ssx \wedge \neg srx)) && (4.8)\\
)
\end{aligned}
$$

The displayed formula describes some properties of a correct run: (4.4) ensures that when S has sent and received an even-coloured message and corresponding acknowledgement, he will next send an odd-coloured message. According to (4.5) this is (socially) guaranteed to lead to a state where R has received this odd-coloured message and where he also acknowledges receipt of it. Then (4.6) says that this will socially lead to a state where S was still sending this odd-coloured message, and has received the corresponding acknowledgement. This will lead then (4.7) to a state where S sends again an even-coloured message that is not yet acknowledged, after which (4.8) at some point, while S was still repeating the even-coloured message, this becomes acknowledged.

Recall that a socially necessary fact φ is an objective formula for which $\langle\langle\rangle\rangle^s_s \Box\varphi$ holds. They express that some unwanted states will never be reached. For instance, we have in our example that $(ssx \wedge sra) \rightarrow (rrx \wedge rsx)$ is a socially necessary fact: if everybody behaves, then if S is sending an even-coloured message which has been acknowledged, then R's last received message must be even-coloured, and his last sent acknowledgement must be odd as well.

To conclude this case study, let us look at some informational properties in the alternating bit protocol, in particular knowledge and social belief. If

the agents cannot assume others behave socially, the agents actually have *no* knowledge whatsoever about the other agent's state. In other words, if agents can act arbitrarily, then very little knowledge ensues. For instance, in $q15$, S knows of course its own state, but has no clue of R's: $q15 \models \neg K_S rrx \wedge \neg K_S \neg rrx \wedge \neg K_S rsa \wedge \neg K_S \neg rsa$.

However, this is not the case for social beliefs, where the agents mutually assume to abide to the social laws. Recall that we have $\mathsf{AltBit} \models \mathrm{SNF}((ssx \wedge sra) \rightarrow (rrx \wedge rsx))$. Hence, in $q15$, if S can assume that everybody behaves socially, he *is* able to infer the state of R, i.e., we have $\mathsf{AltBit}, q15 \models SB_S(rrx \wedge rsx)$: the sender has the social belief about the correct contents of the receiver's local state.

5 Reducing Social ATL to ATL*

In this section we introduce an alternative approach for expressing properties of systems that refer to whether the agents are acting socially or physically. Rather than using our logical language, Social ATEL, introduced in Section 3, we see to what extent we can capture the same notions using only ATL and ATL*. To this end, we introduce the notion of "good" and "bad" states, similar to the "red" and "green" states of Lomuscio and Sergot in [8]. This idea goes in fact back to a discussion in the deontic logic literature regarding the two notions of *ought to do* and *ought to be* (see for instance [10] and the references therein). Although we do not refer to actions in our object language, the framework discussed in the previous section can be seen to fit in the spirit of an *ought to do* approach: in the semantics, we specify which actions are forbidden in which states. In an *ought to be* based deontic logic, the emphasis is on which states should be avoided. A celebrated paper to link the two is [9], where the fact that action α is forbidden, is represented in a dynamic logic like language as $[\alpha]V$, where V is a propositional atom flagging *violation* of a norm. In words: the action α is forbidden iff doing α leads to only states that mark that something bad has happened.

The paper [8] puts this in an interpreted system context [2], where, rather than an atom V, they use a label *green* to signal green states, states that are reached without any violation. Of course the languages of [9], [8] and ours are all different: [9] uses a dynamic deontic logic, [8] interprets a deontic (epistemic) language over runs (structures with a temporal flavour) while we have both ability operators for different norms and temporal operators. A full comparison between the approaches would be an interesting exercise. Here, we will relate SATL to the languages ATL and ATL* that we enrich with a specific atom *good* (which plays the same role as $\neg V$ and *green* above).

More precisely, we modify our Social Action-Based Alternating Transition Systems by introducing an atomic proposition for each agent, which is only true in the current state if the agent arrived here from the previous state using a legal action. So essentially, we are labeling the states based on how the agents arrived at each state. This gives rise to a larger state space in the modified systems, as now we have copies of each state, representing all the combinations of the agents acting socially or physically to reach it. Ideally we would like to be able to reduce properties expressed in Social ATEL into ATL, expressed using these "good" atomic propositions, in order to automatically verify Social ATEL properties using existing ATL model checkers, such as MOCHA. However, since Social ATEL and ATL are interpreted over different structures, it is not feasible to find direct equivalences between the two. We can, however, express interesting properties using ATL and ATL*.

Using the approach outlined above, we look at several types of Social ATEL properties and see how closely we can express these in ATL*. We investigate the relationship between the two using three special cases of G, where G is the coalition of agents cooperating to achieve some objective. We look at the case where G is the grand coalition (Ag), the empty coalition ($\varnothing$), and finally, an arbitrary coalition (G). We show that there is a general pattern between Social ATEL properties and properties expressed in this approach, which holds regardless of the coalition type and the temporal operators being used. Finally, we prove equivalences between general formulae expressed in Social ATEL and formulae expressed using this approach, for each combination of the coalition acting socially or physically, while the other agents act socially or physically.

5.1 Modifying SAATSs

In this section we introduce the atomic propositions used to give a labeling to each state based on how each agent arrived there. This can either be by performing a *legal* action or by simply performing any *physically possible* action. We introduce an atomic proposition, g_i, one for each agent $i \in \mathsf{Ag}$, with the interpretation that g_i is true in the current state if agent i's last action was a legal one. This corresponds to agent i acting in a social manner (for one time step). $GP = \{g_1, \ldots, g_n\}$ is a set of *good* propositions where $GP \subseteq \Phi$. In order to reason about coalitions of agents, we introduce a proposition good(G) which holds if all the agents in G acted socially to reach the current state. good(G) is defined as follows:

$$\text{good}(G) = \bigwedge_{i \in G} g_i$$

Now we must modify SAAETSs with these g_i propositions. It is important to note that the definition of the g_i propositions comes from the $\ell(\alpha)$

function in the given SAAETS. Now, given a SAAETS

$$\mathsf{Sys} = \langle Q, q_0, \mathsf{Ag}, \mathsf{Ac}_1, \ldots, \mathsf{Ac}_n, \rho, \ell, \tau, \Phi, \pi \rangle$$

in order to express properties of systems in this way, we need to convert the system into a modified system as below:

$$\mathsf{Sys}^\circ = \langle Q^\circ, q_0^\circ, \mathsf{Ag}, \mathsf{Ac}_1, \ldots, \mathsf{Ac}_n, \rho^\circ, \tau^\circ, \Phi^\circ, \pi^\circ \rangle$$

where the modified components have the following interpretation:

- Q°: For each state $q \in Q$, we now have at worst $2^{|\mathsf{Ag}|}$ copies of q to represent all the combinations of agents being "good" and "bad" (this is a worst case estimate, since we can leave out states that are not reachable from q_0°). We encode this extra information as $q_{x_1,\ldots,x_n}$, where $x_i \in \{0,1\}$. x_i being 1 means agent i's last action was a social one, whereas a 0 means it was anti-social. The new set of states formed, Q°, in the worst case, is of size $|Q \times 2^{\mathsf{Ag}}|$. See Figures 2 and 3 for an example of how the train system is transformed;

- q_0°: q_0 becomes q_0°, which is an abbreviation of $q^\circ_{0_{x_1,\ldots,x_n}}$, where $\forall i \in \mathsf{Ag}, x_i = 1$. This is the initial state of the system, which is the same as before, except g_i is true for all agents;

- ρ°: $\forall i \in \mathsf{Ag}, \forall \alpha \in \mathsf{Ac}_{\mathsf{Ag}}, \forall q \in Q, \forall x_i \in \{0,1\} : q \in \rho(\alpha) \Leftrightarrow q_{x_1,\ldots,x_n} \in \rho^\circ(\alpha)$;

- Φ°: Φ is updated with the new g_i propositions: $\Phi^\circ = \Phi \cup GP$;

- π°:
$$\forall i \in \mathsf{Ag}, \forall p \in \Phi, \forall q \in Q, \forall x_i \in \{0,1\} : \\ p \in \pi(q) \Leftrightarrow p \in \pi^\circ(q_{x_1,\ldots,x_n})$$
$$\forall i \in \mathsf{Ag}, \forall g_i \in GP : g_i \in \pi^\circ(q_{x_1,\ldots,x_n}) \Leftrightarrow x_i = 1.$$

- τ°: $\forall i \in \mathsf{Ag}, \forall q, q' \in Q, \forall \alpha_i \in \mathsf{Ac}_i, \forall x_i \in \{0,1\} : [\tau(q, \langle \alpha_1, \ldots, \alpha_k \rangle) = q' \Leftrightarrow \tau^\circ(q_{x_1,\ldots,x_n}, \langle \alpha_1, \ldots, \alpha_k \rangle) = q'_{f_1(x_1),\ldots,f_n(x_n)}$, where $f_i(x_i) = 1 \Leftrightarrow q \in \ell(\alpha_i)]$.

So now we have modified SAAETSs to work with these g_i propositions. Firstly, the set of states has been modified. We have a new copy of each state for all combinations of g_i, for all agents. The initial state is the same as before, but g_i is true for all agents, hence the system starts in a good state. The new action precondition function, ρ°, is directly equivalent to ρ for all states and actions, regardless of the g_i propositions. In other words, if an agent can perform α in q, then the agent can perform α in $q_{x_1,\ldots,x_n}$,

no matter what the values of x_i are. The set of atomic propositions, Φ, is updated with the set of good propositions, GP. The truth definition function, π, is the same as before for atomic propositions in Φ. It is updated for the propositions in GP, where a g_i proposition is true in a state where $x_i = 1$. Finally, the transition function, τ, is updated in the following way: transitions are the same as before, but now, if agent i performs a legal action, in the resultant state, g_i will be true and the x_i subscript of the state will be 1. Performing an illegal action results in the g_i proposition being false and the x_i subscript of the state being 0.

Remark 5.1. In this section we do not look at properties that refer to knowledge. We briefly consider how we would modify the epistemic accessibility relations from SAAETSs to account for these good propositions. We propose the following modification to the epistemic accessibility relations:

$$\sim_i^\circ : \forall i \in \mathsf{Ag}, \forall q, q' \in Q, \forall x_i \in \{0,1\} : q \sim_i q' \Leftrightarrow q_{x_1,\ldots,x_n} \sim_i^\circ q'_{x_1,\ldots,x_n}$$

So if two states, q and q' are indistinguishable in Sys, $q_{x_1,\ldots,x_n}$ and $q'_{x_1,\ldots,x_n}$ will be indistinguishable in Sys°. This will now mean that agents will know if they have acted socially, and also if other agents have acted socially. This would allow us to formulate informational properties about systems where agents can reason about the behaviour of the other agents in the system. An alternative way to modify the epistemic accessibility relations would be as follows:

$$\sim_i^\circ : \forall i \in \mathsf{Ag}, \forall q, q' \in Q, \forall x_i, x'_i \in \{0,1\} : q \sim_i q' \Leftrightarrow q_{x_1,\ldots,x_n} \sim_i^\circ q'_{x'_1,\ldots,x'_n}$$

where $x_i = x'_i$.

This is the same as above, but now agents only know whether they have acted socially themselves, and about how other agents have acted is known. This would mean agents would only be able to reason about their own behaviour.

5.2 Reducing some formulae

There are various interesting types of properties we can form with these good(G) propositions. We will construct example properties using the train system Train that we introduced in Example 2.1. Let Train° be the AATS after modifying Train. We can see the affect modifying the systems has by comparing Figure 2, showing the states and transitions of the standard train system, with Figure 3, which shows the same system after being modified. States are assumed to have a reflexive arc labelled 'i, i', which we omit for clarity.

We start by looking at the following Social ATEL formula:

$$\mathsf{Sys}, q \models \langle\langle G \rangle\rangle_s^s \Box \varphi \tag{5.1}$$

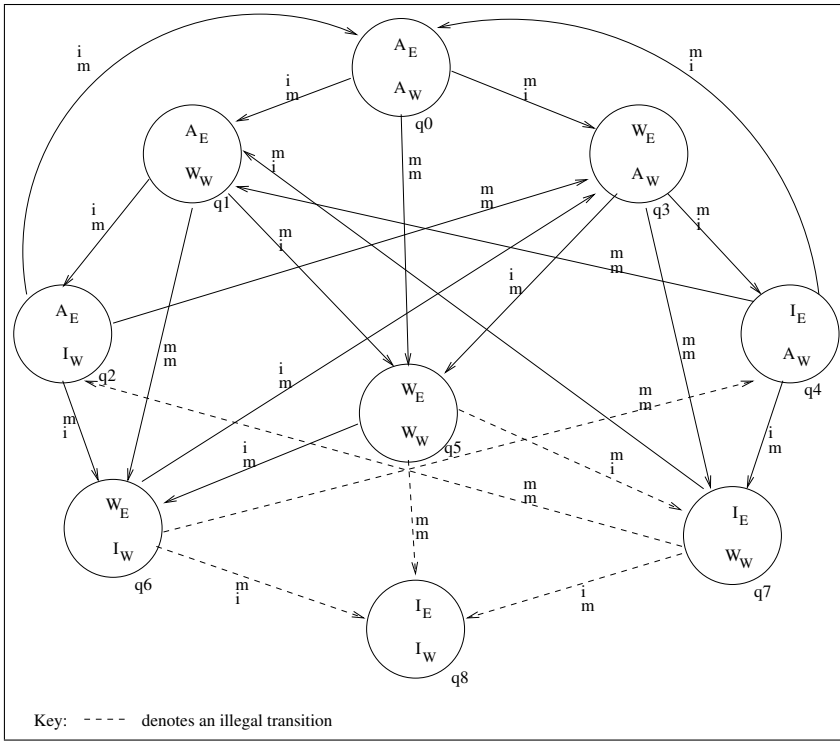


FIGURE 2. State transitions in Train. The top atom in a state refers to train E, as does the top action at each transition.

where φ is assumed to be a propositional logic formula. This Social ATEL formula says that G has a social strategy, such that, no matter what the other agents do, providing they follow social strategies, φ will always be true. We will now try to capture the above using these good(G) propositions where we take G to be the grand coalition, Ag:

$$\mathsf{Sys}^{\circ}, q_{x_1,\ldots,x_n} \models \langle\langle \mathsf{Ag} \rangle\rangle \Box \, (\mathrm{good}(\mathsf{Ag}) \wedge \varphi) \tag{5.2}$$

where $\forall i \in \mathsf{Ag} : x_i = 1$. This formula states that the grand coalition of agents has a strategy such that it will always be the case that good(Ag) is true and φ is true at the same time. This appears to express the same as (5.1) above. If we refer to the train scenario we can formulate the following example property:

$$\mathsf{Sys}^{\circ}, q_{0_{11}} \models \langle\langle \mathsf{Ag} \rangle\rangle \Box \, (\mathrm{good}(\mathsf{Ag}) \wedge \neg \, (\mathsf{in}_E \wedge \mathsf{in}_W)) \tag{5.3}$$

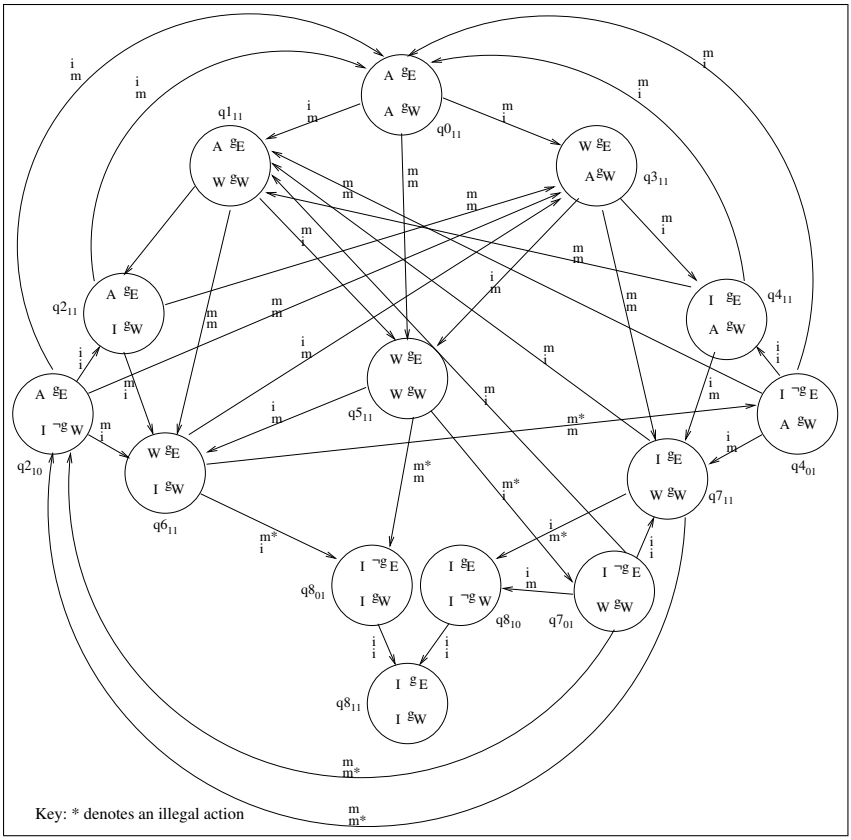


FIGURE 3. State transitions in Train°.

So this property states that the grand coalition should have a strategy so that it is always the case that the agents follow only social strategies and that both of the trains are not in the tunnel at the same time. Property (5.3) passes when we model check it in our implementation of Train°.

Following on from what we said about acting socially leading to the goal, we can formulate the following, where now we take G to be the empty set, $\varnothing$:

$$\mathsf{Sys}^{\circ}, q_{x_1,\ldots,x_n} \models \langle\langle\rangle\rangle \, (\, \Box \, \mathrm{good}(\mathsf{Ag}) \rightarrow \, \Box \varphi) \tag{5.4}$$

where $\forall i \in \mathsf{Ag} : x_i = 1$ and φ is assumed to be a propositional logic formula. The above ATL* formula reads that on all paths, no matter what any of the agents do, good(Ag) always being true implies that φ will always be true. In other words, if the agents *always* follow social strategies, the goal, φ,

is achieved. Referring back to the train scenario, we look at this example property:

$$\mathsf{Sys}^{\circ}, q_{0_{11}} \models \langle\langle\rangle\rangle \left(\Box \, \mathrm{good}(\mathsf{Ag}) \rightarrow \Box \neg (\mathsf{in}_E \wedge \mathsf{in}_W) \right) \tag{5.5}$$

So this reads, no matter what the agents do, on all paths, if the agents always follow social strategies, this implies that the trains will never enter the tunnel at the same time. This sounds intuitively correct based on what we said earlier about social strategies leading to the objective of the social law being satisfied. Using Figure 3, we leave it to the reader to check that (5.5) is satisfied.

Let us now consider

$$\mathsf{Sys}^{\circ}, q_{x_1,\ldots,x_n} \models \langle\langle\rangle\rangle \left(\Box \, \mathrm{good}(\mathsf{Ag}) \rightarrow \Box \varphi \right) \wedge \langle\langle \mathsf{Ag} \rangle\rangle \Box \, \mathrm{good}(\mathsf{Ag}) \tag{5.6}$$

where $\forall i \in \mathsf{Ag} : x_i = 1$ and φ is assumed to be a propositional logic formula. So, as before in (5.4), this formula says that the agents always acting socially implies that the goal will always be achieved, but now also the grand coalition should have a strategy such that they will always act in accordance with the social laws. However, if we refer back to the definition of SAAETSs we see that there is a condition which states that agents should always have at least one legal action available to them in every state of the system. As Sys° is a modified version of Sys, this condition still holds in Sys°, thus guaranteeing that there is at least one social path, hence Ag will *always* have a strategy to ensure good(Ag) is true. This makes this property redundant, as it is directly equivalent to (5.4).

Finally we consider the case where we have an arbitrary coalition, G:

$$\mathsf{Sys}^{\circ}, q_{x_1,\ldots,x_n} \models \langle\langle G \rangle\rangle \left(\Box \, \mathrm{good}(G) \wedge \left(\Box \, \mathrm{good}(\bar{G}) \rightarrow \Box \varphi \right) \right) \tag{5.7}$$

where $\forall i \in \mathsf{Ag} : x_i = 1$ and φ is assumed to be a propositional logic formula. So this formula reads that coalition G has a strategy so that agents in G are always good and that if the other agents in the system are always good then φ will always hold. We needed to separate out the good(Ag) proposition into good(G) and good($\bar{G}$), as G has no control over what the agents in $\bar{G}$ do. Therefore, we can precisely capture (5.1) in this way. If we refer to the trains example, we can formulate a property such as the following:

$$\mathsf{Sys}^{\circ}, q_{0_{11}} \models \langle\langle E \rangle\rangle \left(\Box \, \mathrm{good}(E) \wedge \left(\Box \, \mathrm{good}(W) \rightarrow \Box \neg (\mathsf{in}_E \wedge \mathsf{in}_W) \right) \right) \tag{5.8}$$

This property states that the Eastbound train has a strategy so that it is always social, and if the Westbound train is always social, then the two trains will never be in the tunnel at the same time. This property holds and can be verified by inspection of Figure 3.

Now we consider another type of Social ATEL formula:

$$\mathsf{Sys}, q \models \langle\langle G \rangle\rangle^{s}_{s} \Diamond \varphi \tag{5.9}$$

where φ is assumed to be a propositional logic formula. This formula says that G has a social strategy, such that, no matter what the other agents do, as long as they all only follow social strategies, then φ will be true either now or at some point in the future. We will now try to express a similar property using good(G) propositions, where we take G to be the empty set:

$$\mathsf{Sys}^{\circ}, q_{x_1,\ldots,x_n} \models \langle\langle\rangle\rangle \left(\Box \, \mathrm{good}(\mathsf{Ag}) \rightarrow \Diamond \varphi \right) \tag{5.10}$$

where $\forall i \in \mathsf{Ag} : x_i = 1$ and φ is assumed to be a propositional logic formula. This formula says that no matter what the agents do, on all paths, good(Ag) always being true means that φ will either be true now or at some point in the future. That is to say that if all the agents always act socially, the goal φ will eventually be achieved.

We can also consider situations in which the *other* agents are constrained to social strategies and the coalition of interest can act in an unconstrained manner:

$$\mathsf{Sys}, q \models \langle\langle G \rangle\rangle^{s}_{p} \Box \varphi \tag{5.11}$$

where φ is assumed to be a propositional logic formula. This formula states that G has a strategy to ensure φ is always true, providing the other agents always act in a social manner. We can express something similar to this in the good states approach in the following way:

$$\mathsf{Sys}^{\circ}, q_{x_1,\ldots,x_n} \models \langle\langle G \rangle\rangle \left(\Box \, \mathrm{good}(\bar{G}) \rightarrow \Box \varphi \right) \tag{5.12}$$

where $\forall i \in \mathsf{Ag} : x_i = 1$ and φ is assumed to be a propositional logic formula. So, here we are saying that G has a strategy such that no matter what the other agents in the system do, providing the other agents follow only social strategies, G can always achieve φ. We can look at an example property of the same type as (5.12):

$$\mathsf{Sys}^{\circ}, q_{0_{11}} \models \langle\langle W \rangle\rangle \left(\Box \, \mathrm{good}(E) \rightarrow \Box \neg \left(\mathsf{in}_E \wedge \mathsf{in}_W \right) \right) \tag{5.13}$$

This property states that the westbound train has a strategy so that if the eastbound train always acts socially, then the trains will never enter the tunnel at the same time. This property holds and can be verified by inspection of Figure 3.

5.3 Reducing Social ATEL to ATL*

After investigating the above formulae, we have noticed a general pattern between formulae of Social ATEL and formulae expressed in this good states approach, which seems to follow regardless of the coalition type (grand, empty or arbitrary coalition) and the temporal operator being used.

5.3.1 Bisimulations between computations

Thinking in terms of strategies, there are in general more strategies σ_i° for an agent i in S° than there are strategies σ_i for him in S. To see this consider the following example.

Example 5.2. Suppose we have two agents, called 1 and 2. Suppose agent 1 can perform three actions in q: actions a and b (which we assume are legal) and action c (an illegal action). Suppose agent 2 has two actions at his disposal in q: action d (legal) and e (illegal), and no other action to choose there. Suppose furthermore that the transition function τ is such that $\tau(q, \langle a, d\rangle) = \tau(q, \langle c, d\rangle) = q$. In S°, this gives rise to transitions $\tau(q_{11}\langle a, d\rangle) = q_{11}$, while $\tau(q_{11}, \langle c, d\rangle) = q_{01}$. This enables agent 2 in S° to allow for strategies that depend on how lawful agent 1 behaved in the past in S°. For instance, agent 2 might play a kind of 'tit for tat' by using strategy σ with $\sigma(q_{11}q_{11}) = d$, but at the same time $\sigma(q_{11}q_{01}) = e$: a legal action of agent 1 is replied to by a legal action of agent 2, but after an illegal move of agent 1, agent 2 responds by an illegal action. Notice that such a strategy cannot be implemented in S, since both sequences $q_{11}q_{01}$ and $q_{11}q_{01}$ only correspond to the single sequence qq in S.

However, as we will see, there is a way to connect *computations* in both systems directly to each other.

Definition 5.3. We say that a computation $\lambda = q_0q_1q_2\ldots$ is *compliant* with strategy profile σ_{Ag}, if $\mathsf{comp}(\sigma_{\mathsf{Ag}}, q_0) = \lambda$. That is, the computation can be seen as the effect of applying the strategy profile σ_{Ag} to q_0. If such a strategy profile exists for λ we also say that λ is an *enforceable* computation in the given system.

For any state $q_{x_1\ldots x_n} \in Q^\circ$, let $\mathrm{proj}_Q(q_{x_1\ldots x_n})$ be its corresponding state $q \in Q$. Similarly, for a sequence of states $\vec{s}$ in Q°, the sequence $\mathrm{proj}_{Q^+}(\vec{s})$ denotes the point-wise projection of every state in the sequence to the corresponding state in Q.

Given a system S and its associated system with good states S°, let $\lambda = q_0q_1\ldots$ be a computation in S, and $\lambda^\circ = s_0s_1\ldots$ a computation in S°. We say that λ and λ° *bisimulate* if there are two strategy profiles, σ_{Ag} in S and $\sigma^\circ_{\mathsf{Ag}}$ in S° such that

1. λ is compliant with σ_{Ag} and λ° is compliant with $\sigma^\circ_{\mathsf{Ag}}$
2. for every $u \in \mathbb{N}$, and every $i \in \mathsf{Ag}$, $\sigma_i(q_0\ldots\lambda[u]) = \sigma_i^\circ(s_0,\ldots\lambda^\circ[u])$
3. for every $u \in \mathbb{N}$, $\lambda[u] = \mathrm{proj}_Q(\lambda^\circ[u])$

We say in such a case also that λ and λ° *bisimulate with strategy profiles* σ_{Ag} and $\sigma^\circ_{\mathsf{Ag}}$. Notation: $\langle\lambda, \sigma_{\mathsf{Ag}}\rangle \simeq \langle\lambda^\circ, \sigma^\circ_{\mathsf{Ag}}\rangle$.

Note that a computation need not be compliant with any strategy, and, moreover, if it is compliant with one, it will in general be compliant with several others as well. The latter is so because of two reasons: first of all, a computation only specifies what the transitions are *within* a particular sequence of states, and says nothing about choices on states that do not occur in that computation. Secondly, even within a computation, it is well possible that a transition from q_i to q_{i+1} can be the effect of different choices by the grand coalition Ag. For two computations to be bisimilar, Definition 5.3 demands however that there are two strategies, one for each computation, in which exactly the same actions are taken, at every state in the computation. Moreover, item 3 guarantees that the computations also only visit corresponding states.

Let an *objective temporal formula* ψ be defined as follows:

$$\psi := p \mid \neg\psi \mid \psi \wedge \psi \mid \bigcirc\psi \mid \ \Box\psi \mid \psi \,\mathcal{U}\, \psi$$

Such formulae can be interpreted on infinite paths of states $\lambda = q_0 q_1 \ldots$ in a straightforward way:

$$\begin{array}{lll}
S, q_0 q_1 \ldots \models p & \text{iff} & S, q_0 \models p \\
S, q_0 q_1 \ldots \models \psi_1 \wedge \psi_2 & \text{iff} & S, q_0 q_1 \ldots \models \psi_1 \text{ and } S, q_0 q_1 \ldots \models \psi_2 \\
S, q_0 q_1 \ldots \models \bigcirc\psi & \text{iff} & S, q_1 \ldots \models \psi \\
S, q_0 q_1 \ldots \models \Box\psi & \text{iff} & \forall i, S, q_i q_{i+1} \ldots \models \psi \\
S, q_0 q_1 \ldots \models \psi_1 \,\mathcal{U}\, \psi_2 & \text{iff} & \exists i \text{ s.t. } S, q_i q_{i+1} \ldots \models \psi_2 \text{ and} \\
 & & \quad \forall 0 \leq j < i, S, q_j q_{j+1} \ldots \models \psi_1
\end{array}$$

Note that, in particular, since the propositions g_i are atomic propositions in S°, we can interpret them on an infinite path in S°.

Lemma 5.4. Let $\lambda = q_0 q_1 q_2 \ldots$ and $\lambda^\circ = s_0 s_1 s_2 \ldots$. Suppose furthermore that $\langle \lambda, \sigma_{\mathsf{Ag}} \rangle \simeq \langle \lambda^\circ, \sigma^\circ_{\mathsf{Ag}} \rangle$. Then:

1. for every $u \in \mathbb{N}$ and every objective temporal formula ψ:

$$S, \lambda[u] \models \psi \ \text{ iff } \ S^\circ, \lambda^\circ[u] \models \psi$$

2. for every $u \in \mathbb{N}, i \in \mathsf{Ag}$,

$$\lambda[u] \in \ell(\sigma_i(q_0 \ldots \lambda[u])) \ \text{ iff } \ S^\circ, \lambda^\circ[u+1] \models g_i$$

Proof.

1. Let λ and λ° be as specified. Recall that, by item 3 of Definition 5.3, for all $u \in \mathbb{N}$, $\lambda[u] = \mathrm{proj}_Q(\lambda^\circ[u])$ $(*)$. We now prove by induction

ψ that for all $u \in \mathbb{N}$, $S, \lambda[u] \models \psi$ iff $S^\circ, \lambda^\circ[u] \models \psi$. For atomic propositions p, this follows immediately from the equivalence $(*)$ and the definition of π°. Now suppose the property is proven for ψ: we only do the $\bigcirc$-case. Let $u \in \mathbb{N}$ be arbitrary.

$$\begin{array}{lll} S, \lambda[u] \models \bigcirc\psi & \text{iff} & \text{(definition of } \bigcirc\text{)} \\ S, \lambda[u+1] \models \psi & \text{iff} & \text{(induction hypothesis)} \\ S^\circ, \lambda^\circ[u+1] \models \psi & \text{iff} & \text{(definition of } \bigcirc\text{)} \\ S^\circ, \lambda^\circ[u] \models \bigcirc\psi & & \end{array}$$

2. Let $\lambda[u+1] = q$, for some $q \in Q$. Note that, by item 3 of Definition 5.3 $\text{proj}_Q(\lambda^\circ[u+1]) = q$. So, $\lambda^\circ[u+1]$ is $q_{x_1 \ldots x_{i-1}, x_i, x_{i+1}, \ldots x_n}$ for some sequence $x_1 \ldots x_{i-1}, x_i, x_{i+1}, \ldots x_n \in \{0,1\}^n$. By definition of τ°, we have $x_i = 1$ iff $\lambda[u] \in \ell(\sigma_i(q_0 \ldots \lambda[u]))$. Moreover, by the truth definition of g_i, we have $x_i = 1$ iff $S^\circ, \lambda^\circ \models g_i$ The result now immediately follows from the above two statements.

Q.E.D.

So, if we have two computations that bisimulate, then according to item 1 of Lemma 5.4, they verify the same objective temporal formulae, and, by item 2, a choice for an agent i in the original system is legal, if and only if in the associated system, in the state that results the proposition g_i is true.

Definition 5.5. Let a computation λ be compatible with strategy profile σ_{Ag}. We say that coalition G behaves socially according to this profile along the computation λ, if $\forall u \in \mathbb{N}, \forall i \in G,\ (\lambda[u] \in \ell(\sigma_i(q_0 \ldots \lambda[u])))$.

So, G behaves social along λ, if it has a contribution to generate the computation λ that consists only of social actions.

Corollary 5.6. Suppose that λ and λ° are bisimilar computations, with strategy profiles σ_{Ag} and $\sigma^\circ_{\mathsf{Ag}}$, respectively. Let G be an arbitrary coalition. Then

1. G behaves social according to σ_{Ag} along λ iff $S^\circ, \lambda^\circ \models \Box\,\text{good}(G)$
2. If for all $i \in G$, $\sigma_i \in \Delta_i$, then $S^\circ, \lambda^\circ \models \Box\,\text{good}(G)$.

Proof.

1. Note that item 2 of Lemma 5.4 implies

$$\forall u \in \mathbb{N}, \forall i \in G \quad (\lambda[u] \in \ell(\sigma_i(q_0 \ldots \lambda[u])) \text{ iff } S^\circ, \lambda^\circ[u+1] \models g_i)$$

This, in turn, implies

$$(\forall u \in \mathbb{N}, \forall i \in G \;\; \lambda[u] \in \ell(\sigma_i(q_0 \dots \lambda[u]))) \text{ iff } \forall u \in \mathbb{N}, \forall i \in G(S^\circ, \lambda^\circ[u+1] \models g_i)$$

The left-hand side of the above 'iff' states that G behaves social according to σ_{Ag} along λ, and the right-hand side is equivalent to $S^\circ, \lambda^\circ \models \Box \,\mathrm{good}(G)$.

2. This follows immediately from the above: note that if for all $i \in G$, $\sigma_i \in \Delta_i$, then $\forall u \in \mathbb{N}, \forall i \in G \; \lambda[u] \in \ell(\sigma_i(q_0 \dots \lambda[u]))$.

Q.E.D.

The converse of item 2 of Corollary 5.6 is in general not true: if $S^\circ, \lambda^\circ \models \mathrm{good}(G)$, then we only know that along the computation λ, all members of G behave well, but outside λ, they may not and hence their strategy need not be social.

Definition 5.7. Let λ be a computation. We say that strategies σ_i and σ_i' for i *coincide along* λ, written, $\sigma_i \equiv_\lambda \sigma_i'$, if for all $u \in \mathbb{N}$, $\sigma_i(\lambda[0] \dots \lambda[u]) = \sigma_i'(\lambda[0] \dots \lambda[u])$. For a coalition, $\sigma_G \equiv \sigma_G'$, if for every $i \in G$, $\sigma_i \equiv_\lambda \sigma_i'$. Moreover, for any strategy profile σ_{Ag}, and τ_i a strategy for i, we write $\sigma[\tau_i/\sigma_i]$ for the strategy profile that is like σ_{Ag}, except that for agent i, the component σ_i is replaced by τ_i. Similarly for $\sigma[\tau_G/\sigma_G]$.

It is easy to see that if σ_{Ag} is compliant with λ, and $\sigma_i \equiv_\lambda \sigma_i'$, then $\sigma[\sigma_i'/\sigma_i]_{\mathsf{Ag}}$ is also compliant with λ.

Lemma 5.8. Suppose λ is a computation, and strategy profile σ is compliant with it. Suppose furthermore that G behaves social according to this profile along λ. Then there exists a set of strategies τ_G, such that $\tau_G \in \Delta_G$ and $\sigma_G \equiv_\lambda \tau_G$.

Proof. For every finite prefix $q_0 q_1 \dots q_n$ of λ, and every $i \in G$, take $\tau_i(q_0 q_1 \dots q_n) = \sigma_i(q_0 q_1 \dots q_n)$. For this choice, we obviously have $\sigma_i \equiv_\lambda \tau_i$. Also, note that every action $\tau_i(q_0 q_1 \dots q_n)$ is a legal choice, because i behaves social according to the profile σ along λ. Since by definition of a system S every agent can always perform a social action, we can extend τ_i for any other sequence $\vec{s}$ of states in such a way that the choice $\tau_i(\vec{s})$ is a legal action. It is clear that τ_G satisfies the conditions of the lemma. Q.E.D.

We noted above that in general there are more strategies for a coalition in S° than there are in S. Our main result connecting a system S with S° now says that all enforceable computations in one of the systems have a computation that is bisimilar in the other.

Theorem 5.9. Let S and S° be as defined before. Suppose λ is compliant with profile σ_{Ag}. Then: there is a computation λ° in S° and a strategy profile $\sigma^\circ_{\mathsf{Ag}}$, such that $\langle \lambda, \sigma_{\mathsf{Ag}} \rangle \simeq \langle \lambda^\circ, \sigma^\circ_{\mathsf{Ag}} \rangle$. The converse is also true: for every computation λ° in S° that is compliant with a strategy $\lambda^\circ_{\mathsf{Ag}}$ we can find a strategy profile σ_{Ag} and computation λ in S such that $\langle \lambda, \sigma_{\mathsf{Ag}} \rangle \simeq \langle \lambda^\circ, \sigma^\circ_{\mathsf{Ag}} \rangle$.

Proof. From S to S°: let $\lambda = q_0 q_1 \ldots$ be an enforceable computation in S and let it be compliant with σ_{Ag}. Let σ° be a strategy in S° for agent i satisfying:

$$\sigma^\circ_i(\vec{s}) = \sigma_i(\mathrm{proj}_{Q^+}(\vec{s}))$$

The strategy profile $\sigma^\circ_{\mathsf{Ag}}$ generates a computation $\lambda^\circ = s_0 s_1 \ldots$ for any s_0 with $\mathrm{proj}_Q(s_0) = q_0$. Hence, by definition, $\sigma^\circ_{\mathsf{Ag}}$ is compliant with λ°. This shows item 1 of Definition 5.3. By definition of this λ° and σ°_i, also item 2 of Definition 5.3 is satisfied. We now demonstrate item 3 using induction on the length of the computations, u. If $u = 0$, $\lambda[0] = q_0 = \mathrm{proj}_Q(s_0) = \lambda^\circ[0]$ (note that $q_0 = \mathrm{proj}_Q(s_0)$ by our choice of s_0). Now suppose the following induction hypothesis (ih) holds: $\forall n \leq u : \lambda[n] = \lambda^\circ[n]$. Then

$$\begin{array}{lcl} \lambda[u+1] & = & \text{by definition of computation} \\ \tau(\lambda[u], \sigma_{\mathsf{Ag}}(q_0 \ldots \lambda[u])) & = & \text{definition of } \tau^\circ \text{ and } ih \\ \tau^\circ(\lambda^\circ[u], \sigma^\circ_{\mathsf{Ag}}(s_0 \ldots \lambda^\circ[u])) & = & \text{by definition of computation} \\ \lambda^\circ[u+1] & & \end{array}$$

From S° to S: Let $\lambda^\circ = s_0 s_1 \ldots$. Since λ° is enforceable, we know that λ° is generated by some strategy profile $\sigma^\circ_{\mathsf{Ag}}$. Let $\lambda = \mathrm{proj}_{Q^+}(\lambda^\circ) = \mathrm{proj}_Q(s_0)\mathrm{proj}_Q(s_1)\ldots$. It is easy to see that λ is generated by any strategy profile σ_{Ag} that satisfies, for any $i \in \mathsf{Ag}$:

$$\sigma_i(\mathrm{proj}_Q(s_0)\mathrm{proj}_Q(s_1)\ldots\mathrm{proj}_Q(s_u)) = \sigma^\circ_i(s_0 s_1 \ldots s_u)$$

Hence, σ_{Ag} is compliant with λ. Item 2 of Definition 5.3 follows directly, as does item 3 of that definition. Q.E.D.

Going briefly back to Example 5.2, although the strategy τ in S° for agent 2 that simulates 'tit for tat' has not immediately a counterpart in S, in a *computation* λ°, agent 1 must have made up his mind between behaving 'good' and 'bad', and hence we either have the computation $q_{11}q_{11}\ldots$ or $q_{11}q_{01}\ldots$. And *those do have* a counterpart $qq\ldots$ in S, and they are generated by different strategy profiles: in the first, 1 plays initially a, in the second, it would be c.

5.3.2 Proving some reductions

The following reduction holds where G acts socially and the other agents also act socially.

Proposition 5.10. Let $T\psi$ be an objective path formula and let q° be such that $\text{proj}_Q(q^\circ) = q$.

$$\mathsf{Sys}, q \models \langle\!\langle G \rangle\!\rangle^s_s T\psi \Leftrightarrow \mathsf{Sys}^\circ, q^\circ \models \langle\!\langle G \rangle\!\rangle \big(\Box \,\text{good}(G) \wedge \big(\Box \,\text{good}(\bar{G}) \rightarrow T\psi \big) \big)$$

Proof. Let ψ be an objective path formula. Suppose $\mathsf{Sys}, q \models \langle\!\langle G \rangle\!\rangle^s_s T\psi$. This means that there is a social strategy $\sigma_G \in \Delta_G$ for G, such that for any social strategy $\sigma_{\bar{G}} \in \Delta_{\bar{G}}$ for $\bar{G}$, if $\lambda = \mathsf{comp}(\langle \sigma_G, \sigma_{\bar{G}} \rangle, q)$, then $\mathsf{Sys}, \lambda \models T\psi$. $(*)$.

Now consider an arbitrary state q° in Sys° for which $\text{proj}_Q(q^\circ) = q$. Let, for every $i \in G$, the strategy σ°_i for agent i be *fixed:* it is exactly like σ_i on every projected sequence, that is, let $\sigma^\circ_i(\vec{s}; s)$ be $\sigma_i(\text{proj}_{Q^+}(\vec{s}); \text{proj}_Q(s))$. Now consider an *arbitrary strategy* $\sigma^\circ_{\bar{G}}$ for $\bar{G}$, and let $\lambda^\circ = \mathsf{comp}(\langle \sigma^\circ_G, \sigma^\circ_{\bar{G}} \rangle, q^\circ)$. If we can show that for any such λ°, we have

$$S^\circ, \lambda^\circ \models \big(\Box \,\text{good}(G) \wedge \big(\Box \,\text{good}(\bar{G}) \rightarrow T\psi \big) \big),$$

we are done.

To show that $\mathsf{Sys}^\circ, \lambda^\circ \models \Box \,\text{good}(G)$, recall that σ_G is a social strategy for G, and then apply Corollary 5.6, item 2. To show that also $\mathsf{Sys}^\circ, \lambda^\circ \models \Box \,\text{good}(\bar{G}) \rightarrow T\psi$, assume that $\mathsf{Sys}^\circ, \lambda^\circ \models \Box \,\text{good}(\bar{G})$. Recall that λ° is a computation $\lambda^\circ = \mathsf{comp}(\langle \sigma^\circ_G, \sigma^\circ_{\bar{G}} \rangle, q^\circ)$, where σ°_G is fixed. To stress that the computation depends on the strategy $\bar{\sigma}^\circ_{\bar{G}}$, we will also write $\lambda^\circ(\bar{\sigma}^\circ_{\bar{G}})$ for λ°. Obviously, each such $\lambda^\circ(\sigma^\circ_{\bar{G}})$ is compliant with $\langle \sigma^\circ_G, \sigma^\circ_{\bar{G}} \rangle$. For each such $\sigma^\circ_{\bar{G}}$, let the strategy $\sigma_{\bar{G}}$ in Sys be obtained from $\sigma^\circ_{\bar{G}}$ in the standard way: it has to satisfy, for every agent $\bar{g}$ in the coalition $\bar{G}$, that $\sigma_{\bar{g}}(\text{proj}_Q(\lambda^\circ[0]) \ldots \text{proj}_Q(\lambda^\circ[n])) = \sigma^\circ_{\bar{g}}(\lambda^\circ[0], \ldots \lambda^\circ[n])$. Let $\lambda(\sigma_{\bar{G}}) = \mathsf{comp}(\langle \sigma_G, \sigma_{\bar{G}} \rangle, q)$, one computation for each $\sigma_{\bar{G}}$. It is clear that, for each $\sigma^\circ_{\bar{G}}$, $\lambda(\sigma_{\bar{G}})$ and $\lambda^\circ(\sigma^\circ_{\bar{G}})$ are bisimilar computations, with strategy profiles $\langle \sigma_G, \sigma_{\bar{G}} \rangle$ and $\langle \sigma^\circ_G, \sigma^\circ_{\bar{G}} \rangle$, respectively. Since we assumed $\mathsf{Sys}^\circ, \lambda^\circ(\sigma^\circ_{\bar{G}}) \models \Box \,\text{good}(\bar{G})$, by Corollary 5.6, item 1, we have that $\bar{G}$ behaves social according to $\langle \sigma_G, \sigma_{\bar{G}} \rangle$ along λ. By Lemma 5.8, there is a strategy τ_G for G such that $\tau_G \equiv_\lambda \sigma_G$, and $\tau_G \in \Delta_G$. That is, τ_G is a social strategy. By $(*)$, we then have $\mathsf{Sys}, \lambda \models T\psi$. Since λ and λ° are bisimulating computations, we also have $\mathsf{Sys}^\circ, \lambda^\circ \models T\psi$, which had to be proven.

For the converse, let $\mathsf{Sys}^\circ, q^\circ \models \langle\!\langle G \rangle\!\rangle(\Box \,\text{good}(G) \wedge (\Box \,\text{good}(\bar{G}) \rightarrow T\psi))$. By the semantics of ATL, this means that there is a strategy σ°_G for G, such that for all strategies $\sigma^\circ_{\bar{G}}$, if $\lambda^\circ = \mathsf{comp}(\langle \sigma^\circ_G, \sigma^\circ_{\bar{G}} \rangle, q^\circ)$, then $\mathsf{Sys}^\circ, \lambda^\circ \models (\Box \,\text{good}(G) \wedge (\Box \,\text{good}(\bar{G}) \rightarrow T\psi))$. Note that σ°_G is fixed. From Corollary 2 we know that every strategy $\langle \sigma_G, \sigma_{\bar{G}} \rangle$ that is bisimilar to $\langle \sigma^\circ_G, \sigma^\circ_{\bar{G}} \rangle$ is such that G behaves social according to $\langle \sigma_G, \sigma_{\bar{G}} \rangle$, i.e., $\sigma_G \in \Delta_G$. For any $q \in Q$, let $q_{\vec{1}}$ be the a corresponding state in S° with all indices being a 1. Now we define a strategy σ_i, for each $i \in G$, as follows:

$$\sigma_i(q_0 q_1 \ldots q_n) = \sigma^\circ_i(q_{0_{\vec{1}}} q_{1_{\vec{1}}} \ldots q_{n_{\vec{1}}})$$

That is, σ_i 'copies' the behaviour as prescribed in σ_i° on 'good paths'. We are going to show that σ_G has the property that, if we combine it with any $\sigma_{\bar{G}} \in \Delta_{\bar{G}}$, every computation $\lambda = \mathsf{comp}(\langle \sigma_G, \sigma_{\bar{G}} \rangle, q)$ has the property that $S, \lambda \models T\psi$. This would complete the proof that $S, q \models \langle\langle G \rangle\rangle^s_s T\psi$.

So, take any social strategy $\sigma_{\bar{G}}$ for $\bar{G}$. Take the strategy profile $\sigma_{\mathsf{Ag}} = \langle \sigma_G, \sigma_{\bar{G}} \rangle$. Obviously, this is in Δ_{Ag}, since both coalitions act socially. Let λ be the computation $\mathsf{comp}(\langle \sigma_G, \sigma_{\bar{G}} \rangle, q)$. We know from Theorem 5.9 that there is a computation $\lambda^\bullet$ and a strategy profile $\sigma^\bullet_{\mathsf{Ag}}$ such that $\langle \lambda, \sigma_{\mathsf{Ag}} \rangle \simeq \langle \lambda^\bullet, \sigma^\bullet_{\mathsf{Ag}} \rangle$. Note that we use fresh symbols for this compuation and strategy profile, since λ° and σ° already have a meaning. Now we argue that $\lambda^\bullet$ can be conceived as a computation $\mathsf{comp}(\langle \sigma^\circ_G, \sigma^\circ_{\bar{G}} \rangle, q^\circ)$. First of all, recall that both G and $\bar{G}$ are acting socially in S. Hence, the computation $\lambda^\bullet$ is of the form $q^\circ q_{1_{\bar{1}}} q_{2_{\bar{1}}} \ldots$. In other words, apart from possibly the first state, all states have indices x_i that are all equal to 1! But then, on this computation, we can just assume that the strategy of G is the earlier σ°_G: on sequences with only 1's, we have copied the choices of σ°_G to σ_G and now back to $\lambda^\bullet$. Formally: for all $u \in \mathbb{N}$ and $i \in G$: $\sigma^\bullet_i(q^\circ q_{1_{\bar{1}}} q_{2_{\bar{1}}} \ldots q_{u_{\bar{1}}}) = \sigma_i(qq_1 q_2 \ldots q_u) = \sigma^\circ_i(q^\circ q_{1_{\bar{1}}} q_{2_{\bar{1}}} \ldots q_{u_{\bar{1}}})$. Moreover, since $\sigma_{\mathsf{Ag}} \in \Delta_{\mathsf{Ag}}$, we have $S^\circ, \lambda^\bullet \models \Box(\mathrm{good}(G) \wedge \mathrm{good}(\bar{G}))$. But we know that on such paths, when they are generated by σ_G, that $T\psi$ holds. Now we use Lemma 5.4, to conclude that $S, \lambda \models T\psi$, as required. Q.E.D.

So far we have only looked at the reduction for cases where G is acting socially and the other agents are also acting socially. When we alter the type of strategies that the agents must follow, we obtain the following reductions, of which the proof is similar to that of Proposition 5.10:

Proposition 5.11. Let ψ be an objective path formula, and T a temporal operator. Let q° be such that $\mathrm{proj}_Q(q^\circ) = q$.

1. $\mathsf{Sys}, q \models \langle\langle G \rangle\rangle^p_s T\psi \Leftrightarrow \mathsf{Sys}^\circ, q^\circ \models \langle\langle G \rangle\rangle(\Box\, \mathrm{good}(G) \wedge T\psi)$
2. $\mathsf{Sys}, q \models \langle\langle G \rangle\rangle^s_p T\psi \Leftrightarrow \mathsf{Sys}^\circ, q^\circ \models \langle\langle G \rangle\rangle(\Box\, \mathrm{good}(\bar{G}) \rightarrow T\psi)$
3. $\mathsf{Sys}, q \models \langle\langle G \rangle\rangle^p_p T\psi \Leftrightarrow \mathsf{Sys}^\circ, q^\circ \models \langle\langle G \rangle\rangle T\psi$

There are properties we can express with the good states approach that we can't express with Social ATEL. For example, the following formula:

$$\mathsf{Sys}^\circ, q_{x_1,\ldots,x_n} \models \langle\langle\rangle\rangle\, (\mathrm{good}(\mathsf{Ag})\, \mathcal{U}\, \varphi) \tag{5.14}$$

Comparing this with (5.9) we see that both are generally saying that if Ag act in a social manner then φ will eventually be true, however, (5.14) only requires Ag to follow social strategies *until* φ is satisfied whereas (5.9) requires Ag to *always* act in a social manner.

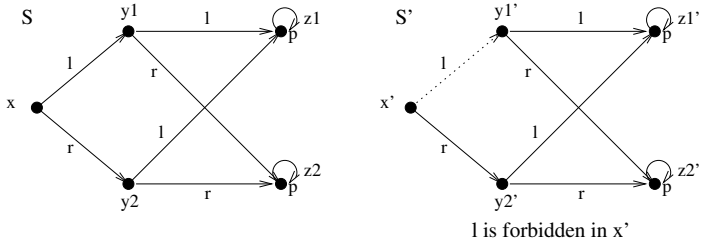


FIGURE 4. Systems S and S'.

We now prove that (5.14) can not be expressed in Social ATEL. Firstly, we introduce two systems, S and S', illustrated in Figure 4. Both of these systems consist of only one agent, which in each state is faced with the choice of two actions: l or r (for left or right, respectively). There is only one atomic proposition, p, which holds in states z_1 and z_2 (z'_1 and z'_2 in S'). Both the systems are the same apart from in S' the action l is forbidden in x'. It does not matter if the agent chooses l or r as both of these actions lead to a state where the same propositions hold. This is given formally by the following lemma:

Lemma 5.12. For the two systems S and S', the following holds:

a. $\forall\varphi \in \Phi : S, y_1 \models \varphi \Leftrightarrow S, y_2 \models \varphi$ and $S, z_1 \models \varphi \Leftrightarrow S, z_2 \models \varphi$;

b. $\forall\varphi \in \Phi : S', y'_1 \models \varphi \Leftrightarrow S', y'_2 \models \varphi$ and $S', z'_1 \models \varphi \Leftrightarrow S', z'_2 \models \varphi$.

Proof. We only prove part a., as the proof of part b. follows exactly the same reasoning. Firstly, we argue $\forall\varphi \in \Phi : S, z_1 \models \varphi \Leftrightarrow S, z_2 \models \varphi$, and we do this using induction on φ. For atomic propositions p, negation and conjunctions, this is straightforward. For the $\langle\langle G\rangle\rangle^p_p T\varphi$ case, where T is an arbitrary temporal operator, suppose the following induction hypothesis holds: $S, z_1 \models \varphi \Leftrightarrow S, z_2 \models \varphi$. To evaluate $S, z_1 \models \langle\langle G\rangle\rangle^p_p T\varphi$ we will consider computations that only visit z_1, and, similarly, for $S, z_2 \models \langle\langle G\rangle\rangle^p_p T\varphi$ gives rise to computations which only visit z_2. So, using induction, the computations visit 'equivalent' states, and we are done.

Finally, we want to prove:

$$\forall\varphi \in \Phi : S, y_1 \models \varphi \Leftrightarrow S, y_2 \models \varphi$$

by induction on φ. The cases of p, $\neg\varphi$ and $\varphi_1 \wedge \varphi_2$ are again straightforward.

For the $\langle\langle G\rangle\rangle^p_p T\varphi$ case, notice how $\tau(y_1, l) = z_1$ and $\tau(y_2, l) = z_1$, and also $\tau(y_1, r) = z_2$ and $\tau(y_2, r) = z_2$. So l and r lead to the same states no matter whether they are performed in y_1 or y_2. As we have already proven

$\forall \varphi \in \Phi : S, z_1 \models \varphi \Leftrightarrow S, z_2 \models \varphi$, it follows that $S, y_1 \models \langle\langle G \rangle\rangle_p^p T\varphi \Leftrightarrow S, y_2 \models \langle\langle G \rangle\rangle_p^p T\varphi$. Q.E.D.

The following theorem illustrates the fact that (5.14) can not be expressed in Social ATEL:

Theorem 5.13. Given the two systems, S and S', shown in Figure 4, we claim that $\forall \varphi \in \Phi$, where Φ is the set of all Social ATEL formulae: $S, x \models \varphi \Leftrightarrow S', x' \models \varphi$, but in the system with good states, $S, x \models \langle\langle\rangle\rangle(\text{good}(\mathsf{Ag}) \,\mathcal{U}\, p)$ while $S', x' \not\models \langle\langle\rangle\rangle(\text{good}(\mathsf{Ag}) \,\mathcal{U}\, p)$.

Proof. We prove $\forall q \in Q, \forall \varphi \in \Phi : S, q \models \varphi \Leftrightarrow S', q' \models \varphi$ by induction on φ. For $q \in \{y_1, y_2, z_1, z_2\}$ this is clear: every q and q' satisfy the same atomic propositions, and furthermore, the transitions lead to 'similar' states.

Let us now look at $q = x \in Q$. The key now is that every physical computation in S starting in x has a corresponding physical computation in S' starting in x', and the same for every legal computation in S: if our agent chooses l in x in S, the next state is y_1, which, by Lemma 5.12 is equivalent to y_2, which, by the above, is equivalent to y_1'. Also, every physical (legal) computation in S' from x' has a corresponding physical (legal) computation in S from x.

Finally, we show that in the system with good states, $S, x \models \langle\langle\rangle\rangle(\text{good}(\mathsf{Ag}) \,\mathcal{U}\, p)$ while $S', x' \not\models \langle\langle\rangle\rangle(\text{good}(\mathsf{Ag}) \,\mathcal{U}\, p)$. Recall we only have one agent. In the good-state system associated with S, good(1) is true everywhere, and we also have $S, x \models \langle\langle\rangle\rangle(\text{good}(\mathsf{Ag}) \,\mathcal{U}\, p)$: along every path, good(1) is true until a p-state is reached. But in S', good(1) is false in $y1'$ and there is a computation $x', y1', z1', z1', \ldots$ along which $\text{good}(1) \,\mathcal{U}\, p$ is false. Q.E.D.

6 Conclusion

In this paper we have extended our social laws framework further. In previous work, the basic framework was extended to incorporate epistemic notions, but here, we have extended this by removing the assumption that all agents in the system will adhere to the social laws. Firstly, we extended the semantic structures to allow us to model physical action pre-conditions and legal action pre-conditions, in turn, allowing us to construct both physical and legal strategies for the agents. With the semantic constructs in place, we introduced our logical language for reasoning about such systems—Social ATEL—based on ATEL but extended to allow us to refer to the type of strategies being followed by the coalition of agents and the other agents in the system.

Our paper is in fact only a modest first hint as to what can be done when allowing for social and anti-social behaviour. There are many venues

to further explore. For instance, our account of knowledge and belief here only suggests some of its use, a deep analysis is not provided here. Especially the notion of social belief deserves further attention. Next, connecting our setup to ATL and ATL* as done in Section 5 left many questions unanswered: we did not provide a precise fit. Finally, there are many other ways one can deal with anti-social behaviour. A more quantitative way of thinking for instance is provided in [3], where a property holds in a *robust* way if it is guaranteed even if the system does at most one non-allowed transition. One might also take a more qualitative view and consider some forms of non-social behaviour more acceptable than others. Finally, to have a norm as a condition on *strategies* rather than *actions* seems an interesting venue to explore further.

References

[1] E.H. Durfee. Distributed problem solving and planning. In G. Weiß, ed., *Multiagent Systems*, pp. 121–164. The MIT Press: Cambridge, MA, 1999.

[2] R. Fagin, J.Y. Halpern, Y. Moses & M.Y. Vardi. *Reasoning About Knowledge*. The MIT Press: Cambridge, MA, 1995.

[3] T. French, C. McCabe-Dansted & M. Reynolds. A temporal logic of robustness. In B. Konev & F. Wolter, eds., *Frontiers of Combining Systems, 6th International Symposium, FroCoS 2007, Liverpool, UK, September 10–12, 2007, Proceedings*, vol. 4720 of *Lecture Notes in Computer Science*, pp. 193–205. Springer, 2007.

[4] W. van der Hoek, M. Roberts & M. Wooldridge. Knowledge and social laws. In F. Dignum, V. Dignum, S. Koenig, S. Kraus, M.P. Singh & M. Wooldridge, eds., *4rd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2005), July 25–29, 2005, Utrecht, The Netherlands*, pp. 674–681. ACM, 2005.

[5] W. van der Hoek, M. Roberts & M. Wooldridge. Social laws in alternating time: Effectiveness, feasibility, and synthesis. *Synthese*, 156(1):1–19, 2007.

[6] W. van der Hoek & M. Wooldridge. Model checking cooperation, knowledge, and time — a case study. *Research in Economics*, 57(3):235–265, 2003.

[7] W. van der Hoek & M. Wooldridge. Time, knowledge, and cooperation: Alternating-time temporal epistemic logic and its applications. *Studia Logica*, 75(1):125–157, 2003.

[8] A. Lomuscio & M. Sergot. Deontic interpreted systems. *Studia Logica*, 75(1):63–92, 2003.

[9] J.-J.Ch. Meyer. A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic*, 29(1):109–136, 1988.

[10] J.-J.Ch. Meyer, R. Wieringa & F. Dignum. The role of deontic logic in the specification of information systems. In J. Chomicki & G. Saake, eds., *Logics for Databases and Information Systems*, pp. 71–115. Kluwer Academic Publishers, 1998.

[11] Y. Moses & M. Tennenholtz. Artificial social systems. *Computers and AI*, 14(6):533–562, 1995.

[12] Y. Shoham & M. Tennenholtz. On the synthesis of useful social laws for artificial agent societies. In W.R. Swartout, ed., *Proceedings of the 10th National Conference on Artificial Intelligence (AAAI-92). San Jose, CA, July 12–16, 1992*. The AAAI Press, 1992.

[13] Y. Shoham & M. Tennenholtz. On social laws for artificial agent societies: Off-line design. In P.E. Agre & S.J. Rosenschein, eds., *Computational Theories of Interaction and Agency*, pp. 597–618. The MIT Press: Cambridge, MA, 1996.

[14] Y. Shoham & M. Tennenholtz. On the emergence of social conventions: Modelling, analysis, and simulations. *Artificial Intelligence*, 94(1–2):139–166, Jul. 1997.

[15] M. Wooldridge. *An Introduction to Multiagent Systems*. John Wiley & Sons, 2002.

A Method for Reasoning about Other Agents' Beliefs from Observations

Alexander Nittka[1]
Richard Booth[2]

[1] Institut für Informatik
Universität Leipzig
Johannisgasse 26
04103 Leipzig, Germany

[2] Faculty of Informatics
Mahasarakham University
Kantarawichai
Mahasarakham 44150, Thailand

nittka@informatik.uni-leipzig.de, richard.b@msu.ac.th

Abstract

This paper is concerned with the problem of how to make inferences about an agent's beliefs based on an observation of how that agent responded to a sequence of revision inputs over time. We collect and review some earlier results for the case where the observation is *complete* in the sense that *(i)* the logical content of all formulae appearing in the observation is known, and *(ii) all* revision inputs received by the agent during the observed period are recorded in the observation. Then we provide new results for the more general case where information in the observation might be distorted due to noise or some revision inputs are missing altogether. Our results are based on the assumption that the agent employs a specific, but plausible, belief revision framework when incorporating new information.

1 Introduction

1.1 Motivation

One of the overall goals of AI research is designing autonomous intelligent agents that are capable of acting successfully in dynamic environments. These environments may be artificial or even natural. In any case, it is very likely that they are "inhabited" by more than one agent. So, an agent will in general have to interact with (some of) the others. On the one hand, the agent—if it does not want to be purely reactive—needs a model of its environment in order to make informed choices of actions that change it in a way that brings the agent closer to achieving its goal. On the other, it

Giacomo Bonanno, Wiebe van der Hoek, Michael Wooldridge (eds.). *Logic and the Foundations of Game and Decision Theory (LOFT 7).* Texts in Logic and Games 3, Amsterdam

also needs to model the other agents, making successful interaction more likely.

Much research has been done on formalising and reasoning about the effects of actions on an environment. Research on an agent's view of the world usually focuses on a first person perspective. How should the agent adapt its beliefs about the world in the light of new information? However, reasoning about *other agents'* beliefs or background knowledge is just as important. This work is intended to contribute to this latter question.

We will adopt a much narrower perspective than reasoning about other agents in their full complexity which includes goals, intentions, (higher order) beliefs, preferences, etc. and restrict our attention to their (propositional) beliefs about the world. We will also forget about the dynamic environment and assume a static world. That is, we will work in a very traditional belief revision setting. But rather than answering the question of how an agent *should* rationally change its beliefs in the light of new information, we address the question of what we can say about an agent we observe in a belief change process.

In [10], the authors use observable actions to draw conclusions about other agents' mental attitudes. But the beliefs of an agent manifest themselves not only in its actions. They may also be observed more directly, e.g., in communication. So indirectly we have access to parts of other agents' belief revision processes. Information they receive is their revision input, responses to that information are a partial description of their beliefs after the revision. From this information we may want to reason about the observed agent. Consider the following scenarios.

- We are directly communicating with another agent, i.e., we are the source of revision inputs for that agent. The feedback provided by the agent will not reflect its entire set of beliefs. To get a more complete picture we may want to infer what else was believed by the agent, what its *background knowledge* might be.

- We observe a dialogue between two or more agents. Beliefs one agent expresses are revision inputs for the others. Due to noise, private messages etc., we might not have access to the entire dialogue—possibly missing some inputs completely. So we have to deal with partial information about the revision inputs.[1] As we might have to deal with the observed agents later, forming a picture of them will be useful.

The information at our disposal for reasoning about another agent $\mathcal{A}$ will be of the following form. We are given a (possibly incomplete) sequence of

[1] This is of course possible in the first case, as well. The communication might take place in several sessions and we do not know which inputs the agent received in between.

(partially known) revision inputs that were received by $\mathcal{A}$. Further we are given information on what the agent believed and did not believe after having received each input. All this information constitutes an observation of the agent. First we will briefly recall results for the case where observations are complete with respect to the revision inputs received by $\mathcal{A}$. These are then used for dealing with the more general case.

The general approach to reasoning about an agent based on observations will be as follows. We assume $\mathcal{A}$ to employ a particular belief revision framework for incorporating revision inputs. We will then try to find a possible initial state of $\mathcal{A}$ that best explains the observation. By initial state we mean $\mathcal{A}$'s epistemic state at the time the observation started. As we do not know the true initial state, we will have to select a reasonable one. This state explains the observation if it yields the beliefs and non-beliefs recorded in the observation given the revision inputs received by the agent. The meaning of *best* in this context will be explained later. The initial state, which can be interpreted as $\mathcal{A}$'s background knowledge, will allow us to reason about beliefs not recorded in the observation.

Many approaches for reasoning about action, belief revision, etc. assume the initial belief state being given and deal with the case of progression through sequences of actions/revision inputs. They say little or nothing about the case where the initial state is not known. In particular with respect to the belief revision literature this work is intended to be a step towards filling this gap.

1.2 Simplifying assumptions

We make several simplifying assumptions which will naturally limit the applicability of the methods developed in this work but at the same time allow a focused analysis of the problem we approach.

As mentioned above, we assume a static world in the sense that the revision inputs and the information about the agent's beliefs refer to the same world. However, it is essential for our work that the revision inputs were received over time. One central point is to exploit having intermediate steps at our disposal. The observed agent itself may only be interested in the final picture of the world. We in contrast want to extract information about the agent from the process of its arriving there.

We restrict ourselves to propositional logic, and all components of an observation are already provided in propositional logic generated from a finite language. That is, we assume that revision inputs, beliefs and non-beliefs are (and are directly observed as) propositional formulae. Agents are assumed to be sincere, i.e., they are not deceptive about their beliefs, although the information may be partial. The observed agent will be referred to as $\mathcal{A}$. We will disregard concepts like (preferences for) sources, competence, context, etc. $\mathcal{A}$ will be assumed to employ a particular belief revision

framework which we describe in detail in Section 2. The only thing that happens during the time of observation is that $\mathcal{A}$ incorporates the revision inputs. In particular, it does not change its revision strategy or learns in any other way. In that sense, we consider the observations to be short term.

We do not investigate *strategies* for extracting as much information about $\mathcal{A}$ as possible. The observing agent simply uses the information provided to reason along the way, being passive in that sense. That is, our focus is not on the *elicitation* of information about other agents; the question of optimising the reasoning process by putting agents in a setting where observations yield the most precise results is another interesting topic which we do not pursue.

From the choice of revision framework it will become apparent that we equate recency with reliability of the information. We are well aware that this is highly debatable. We will briefly address this issue in the conclusion.

For real world applications many of these assumptions have to be dropped or weakened. Many of the issues we disregarded will have to be taken into account. But for the moment we try to keep the number of free variables low in order to give more precise formal results. We hope to convince the reader that even in this very restricted setting we will be able to draw interesting, non-trivial conclusions. Also, we will show that even if these assumptions are correct, there are very strict limitations to what we can *safely* conclude about $\mathcal{A}$.

1.3 Preliminaries

As stated above, the observed agent will be denoted by $\mathcal{A}$. L will be used to denote a propositional language constructed from a finite set of propositional variables $p, q, r, \ldots$, the connectives $\wedge, \vee, \neg, \rightarrow, \leftrightarrow$ and the symbols $\bot$ for some contradiction and $\top$ for some tautology. $\alpha, \beta, \delta, \theta, \lambda, \varphi, \phi, \psi,$ and $\blacktriangle$ (often with subscript) will denote propositional formulae, i.e., particular elements of L. In Section 3, χ will be used as placeholder for an unknown formula. $\vdash$ is the classical entailment relation between a set of formulae and a formula, where we abbreviate $\{\alpha\} \vdash \beta$ by $\alpha \vdash \beta$ for singleton sets. $\mathsf{Cn}(S)$ denotes the set of all logical consequences of a set of formulae S, i.e., $\mathsf{Cn}(S) = \{\alpha \mid S \vdash \alpha\}$.

The revision operation $*$ introduced will be left associative and consequently $K * \varphi_1 * \varphi_2$ is intended to mean $(K * \varphi_1) * \varphi_2$. σ and ρ are used to denote sequences of formulae, () being the empty sequence. The function $\cdot$ denotes concatenation, so $\sigma \cdot \rho$ and $\sigma \cdot \alpha$ represents sequence concatenation and appending a formula to a sequence, respectively.

The structure of the paper will be as follows. Section 2 will introduce the assumed agent model as well as the formal definition of an observation. It further recalls the central results for the case where all revision inputs received by $\mathcal{A}$ during the time of observation are completely known, i.e., in particular the method for calculating the best explaining initial state and its

properties. The section thus summarises [5, 6, 7]. It extends these papers by also discussing the question of how safe conclusions we draw about $\mathcal{A}$ are. Section 3 uses these results to deal with the case where the observation is allowed to be more partial. In particular, some inputs may not have been recorded in the observation (see also [23]) and the logical content of parts of the observation may only be partially known. We show how this lack of information can be represented and dealt with. This paper is intended to give a broad overview over our proposed method for reasoning about an observed agent. Hence, we give only short proofs sketches. Full proofs are available in the first author's PhD thesis [24].

2 Belief Revision Framework, Observation and Explanation

2.1 The assumed belief revision framework

We already mentioned that we will assume the agent to employ a particular belief revision framework. The first thing we will do is describe it. As we consider observations of $\mathcal{A}$'s belief revision behaviour over time, it is obvious that such a framework needs to support iterated revision [12, 17, 21]. Further, an observation may imply that a revision input was in fact not accepted. For example it might be explicitly recorded that after being informed that Manchester is the home of the Beatles, the agent does not believe this statement. Consequently, the assumed revision framework should also account for non-prioritised revision [16, 19], i.e., revision where the input is not necessarily believed after revising.

We will assume $\mathcal{A}$ to employ a belief revision framework [3] that is conceptually similar to the approaches in [4, 9, 20, 25] but is able to handle non-prioritised revision as well. The agent's epistemic state $[\rho, \blacktriangle]$ is made up of two components: *(i)* a sequence ρ of formulae and *(ii)* a single formula $\blacktriangle$, all formulae being elements of L. $\blacktriangle$ stands for the agent's set of core beliefs—the beliefs of the agent it considers "untouchable". One main effect of the core belief is that revision inputs contradicting it will not be accepted into the belief set. ρ is a record of the agent's revision history. Revision by a formula is carried out by simply appending it to ρ. The agent's full set of beliefs $\mathsf{Bel}([\rho, \blacktriangle])$ in the state $[\rho, \blacktriangle]$ is then determined by a particular calculation on ρ and $\blacktriangle$ which uses the function f which maps a sequence σ of propositional formulae to a formula. This is done by starting off with the last element of σ and then going backwards through the sequence collecting those formulae that can be consistently added and forgetting about the remaining ones.

Definition 2.1.

$$f(\beta_k,\dots,\beta_1)=\begin{cases}\beta_1 & k=1\\ \beta_k\wedge f(\beta_{k-1},\dots,\beta_1) & k>1\,\&\,\beta_k\wedge f(\beta_{k-1},\dots,\beta_1)\nvdash\bot\\ f(\beta_{k-1},\dots,\beta_1) & \text{otherwise}\end{cases}$$

As hinted at above, iterated revision is handled quite naturally by the framework. All revision steps are simply recorded and the problem of what $\mathcal{A}$ is to believe after each revision step, in particular whether the input just received is accepted, i.e., is believed, is deferred to the calculation of the beliefs in an epistemic state. In order to calculate them the agent starts with its core belief $\blacktriangle$ and then goes backwards through ρ, adding a formula as an additional conjunct if the resulting formula is consistent. If it is not, then the formula is simply ignored and the next element of ρ is considered. The belief set of $\mathcal{A}$ then is the set of logical consequences of the formula thus constructed.

Definition 2.2. The revision operator $*$ is defined for any epistemic state $[\rho,\blacktriangle]$ and formula φ by setting $[\rho,\blacktriangle]*\varphi=[\rho\cdot\varphi,\blacktriangle]$. The belief set $\mathsf{Bel}([\rho,\blacktriangle])$ in any epistemic state $[\rho,\blacktriangle]$ is $\mathsf{Bel}([\rho,\blacktriangle])=\mathsf{Cn}(f(\rho\cdot\blacktriangle))$.

Note, that we do not prohibit the core belief $\blacktriangle$ to be inconsistent in which case $\mathcal{A}$'s belief set is inconsistent. This is the essential difference of to the linear base-revision operator in [22]. From the definition, it is easy to see that $\mathsf{Bel}([\rho,\blacktriangle])$ is inconsistent if and only if $\blacktriangle$ is inconsistent.

Example 2.3. Consider the epistemic state $[(),\neg p]$ of an agent. The beliefs of the agent in this state are $\mathsf{Cn}(f(\neg p))=\mathsf{Cn}(\neg p)$. If q is received as a new input, we get $[(),\neg p]*q=[(q),\neg p]$ as the new epistemic state. The corresponding beliefs are $\mathsf{Cn}(f(q,\neg p))=\mathsf{Cn}(q\wedge\neg p)$.

A further input $q\to p$ changes the epistemic state to $[(q,q\to p),\neg p]$. Note, that $f(q,q\to p,\neg p)=(q\to p)\wedge\neg p$ and q cannot be consistently added, so now the agent believes the logical consequences of $\neg q\wedge\neg p$.

The revision input p changes the epistemic state to $[(q,q\to p,p),\neg p]$ but the beliefs remain unchanged, as p contradicts the core belief.

Given the state $[\rho,\blacktriangle]$ of $\mathcal{A}$ and a sequence $(\varphi_1,\dots,\varphi_n)$ of revision inputs received in that state we can define the *belief trace* of the agent. This is a sequence of formulae characterising the beliefs of $\mathcal{A}$ after having received each of the inputs starting with the beliefs in $[\rho,\blacktriangle]$.

Definition 2.4. Given a sequence $(\varphi_1,\dots,\varphi_n)$ the *belief trace* $(\mathsf{Bel}_0^\rho,\mathsf{Bel}_1^\rho,\dots,\mathsf{Bel}_n^\rho)$ of an epistemic state $[\rho,\blacktriangle]$ is the sequence of formulae $\mathsf{Bel}_0^\rho=f(\rho\cdot\blacktriangle)$ and $\mathsf{Bel}_i^\rho=f(\rho\cdot(\varphi_1,\dots,\varphi_i,\blacktriangle))$, $1\le i\le n$.

The belief trace in the above example is $(\neg p,q\wedge\neg p,\neg q\wedge\neg p,\neg q\wedge\neg p)$.

2.2 Observations

After having formalised the assumptions about any observed agent, we now turn to the specific information we receive about a particular agent $\mathcal{A}$—some observation on its belief revision behaviour. An observation contains information about revision inputs $\mathcal{A}$ received, what it believed and did not believe upon receiving them.

Definition 2.5. An *observation* $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ is a sequence of triples $(\varphi_i, \theta_i, D_i)$, where for all $1 \leq i \leq n$: φ_i, θ_i, and all $\delta \in D_i$ (D_i is finite) are elements of a finitely generated propositional language L.

The intuitive interpretation of an observation is as follows. After having received the revision inputs φ_1 up to φ_i starting in some initial epistemic state, $\mathcal{A}$ believed at least θ_i but did not believe any element of D_i. In this section, we assume that during the time of the observation $\mathcal{A}$ received exactly the revision inputs recorded in o, in particular we assume that no input was received between φ_i and φ_{i+1}, the observation being correct and complete in that sense. For the θ_i and D_i we assume the observation to be correct but possibly partial, i.e., the agent did indeed believe θ_i and did not believe any $\delta \in D_i$, but there may be formulae ψ for which nothing is known. In this case we have both $\theta_i \nvdash \psi$ and $\psi \nvdash \delta$ for any $\delta \in D_i$. Note that complete ignorance about what the agent believed after a certain revision step can be represented by $\theta_i = \top$ and complete ignorance about what was not believed by $D_i = \varnothing$.

The observation does not necessarily give away explicitly whether a revision input was actually accepted into $\mathcal{A}$'s belief set or not. If $\theta_i \vdash \varphi_i$ then the revision input φ_i must have been accepted and if $\theta_i \vdash \neg\varphi_i$ or $\varphi_i \vdash \delta$ for some $\delta \in D_i$ then it must have been rejected. But if none of these conditions hold, it is not obvious whether an input has been accepted or rejected. Often, none of these two cases can be excluded. One of the aims of our investigation is to draw more precise conclusions with respect to this question.

A given observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ covers only a certain length of time of the agent's revision history. When the observation started, $\mathcal{A}$ already was in some epistemic state $[\rho, \blacktriangle]$. We will give the formal conditions for an initial state to explain an observation o. The intuitive interpretation of o is formally captured by the system of relations in the second condition of the definition.

Definition 2.6. Let $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$. Then $[\rho, \blacktriangle]$ *explains* o (or *is an explanation for* o) if and only if the following two conditions hold.

1. $\blacktriangle \nvdash \bot$
2. For all i such that $1 \leq i \leq n$:

$$\mathsf{Bel}([\rho, \blacktriangle] * \varphi_1 * \cdots * \varphi_i) \vdash \theta_i$$

and

$$\forall \delta \in D_i : \mathsf{Bel}([\rho, \blacktriangle] * \varphi_1 * \cdots * \varphi_i) \nvdash \delta$$

We say $\blacktriangle$ is an *o-acceptable core* iff $[\rho, \blacktriangle]$ explains o for *some* ρ.

For us, an explanation of a given observation o is an epistemic state that verifies the information in o and has a consistent core belief. It is (conceptually) easy to check whether an epistemic state $[\rho, \blacktriangle]$ is an explanation for o. It suffices to confirm that the conditions in Definition 2.6 are satisfied, i.e., that $\blacktriangle$ is consistent and that for all i we have $f(\rho \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle)) \vdash \theta_i$ and $f(\rho \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle)) \nvdash \delta$ for all $\delta \in D_i$. A state with an inconsistent core belief satisfies the second condition if and only if $D_i = \varnothing$ for all i, so there are observations that *could* be explained by such a state. However, we do not consider claiming the agent to be inconsistent worthy of being called an explanation.

Example 2.7. Let $o = \langle (p, q, \varnothing), (q, r, \varnothing) \rangle$ which states that $\mathcal{A}$ after receiving p believes q and after then receiving q believes r. It does not inform us about any non-beliefs of the agent.

$[\rho, \blacktriangle] = [(p \rightarrow q), r]$ explains o because $f(p \rightarrow q, p, r)$ entails q and $f(p \rightarrow q, p, q, r)$ entails r (both are equivalent to $p \wedge q \wedge r$). $[(p \rightarrow q), \top]$ does not explain o because $f(p \rightarrow q, p, q, \top) \equiv p \wedge q \nvdash r$. $[(), p \wedge q \wedge r]$, $[(p \rightarrow q \wedge r), \top]$, $[(\neg p, q, r), s]$, and $[(q \wedge r), \neg p]$ are some further possible explanations for o.

There is never a unique explanation for o, in fact there are infinitely many in case o can be explained. This is why our proposed method for reasoning about $\mathcal{A}$ is to choose one explanation $[\rho, \blacktriangle]$. Using $\blacktriangle$ and the belief trace we then draw our conclusions as follows. Revision inputs consistent with $\blacktriangle$ will be accepted by $\mathcal{A}$, those inconsistent with $\blacktriangle$ are rejected. $\mathcal{A}$'s beliefs after receiving the ith input are characterised by Bel_i^ρ. In Section 2.4 we will discuss the quality of these conclusions and present a method for improving them. But first we have to say how to actually choose one explanation.

2.3 The rational explanation

This section recalls the essential results from [5, 6, 7] for identifying and justifying the best of all possible explanations. A very important property

of the framework is that $\mathcal{A}$'s beliefs after *several* revision steps starting in an initial state can equivalently be expressed as the beliefs after a *single* revision on the same initial state. Intuitively, the agent merges its core belief and all revision inputs received using f into a single formula and then conditions its epistemic state using it.

Proposition 2.8. $\mathsf{Bel}([\rho, \blacktriangle] * \varphi_1 * \cdots * \varphi_i) = \mathsf{Bel}([\rho, \blacktriangle] * f(\varphi_1, \ldots, \varphi_i, \blacktriangle))$.

Proof (Sketch). Note that by Definition 2.4 it suffices to show that

$$f(\rho \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle)) \equiv f(\rho \cdot (f(\varphi_1, \ldots, \varphi_i, \blacktriangle), \blacktriangle)).$$

One property of f that follows from its recursive definition is $f(\sigma \cdot \sigma') \equiv f(\sigma \cdot f(\sigma'))$. If we can show that $f(\varphi_1, \ldots, \varphi_i, \blacktriangle) \equiv f(f(\varphi_1, \ldots, \varphi_i, \blacktriangle), \blacktriangle)$ we are done as then in both cases equivalent formulae have been collected before processing ρ. We can restrict our attention to consistent $\blacktriangle$, in which case $f(\varphi_1, \ldots, \varphi_i, \blacktriangle)$ is consistent and entails $\blacktriangle$. Hence

$$f(f(\varphi_1, \ldots, \varphi_i, \blacktriangle), \blacktriangle) = f(\varphi_1, \ldots, \varphi_i, \blacktriangle) \wedge \blacktriangle \equiv f(\varphi_1, \ldots, \varphi_i, \blacktriangle).$$

Q.E.D.

How does that help to reason about the observed agent $\mathcal{A}$? Recall that an observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$ expresses the information that "revision by φ_1 in the initial state leads to a *new* state (where θ_1 but no element of D_1 is believed) in which revision by φ_2 leads to..." That is, the observation contains bits of information concerning beliefs and non-beliefs in *different* (if related) epistemic states. This proposition now allows us to translate the observation into information about a single state—the initial epistemic state we are after. Note however, that $\blacktriangle$ needs to be known for applying the proposition as otherwise $f(\varphi_1, \ldots, \varphi_i, \blacktriangle)$ cannot be calculated. So, given a core belief $\blacktriangle$, o yields that $\mathcal{A}$ would believe θ_i (and would not believe any $\delta \in D_i$) in case it revised its initial epistemic state by $f(\varphi_1, \ldots, \varphi_i, \blacktriangle)$. This is nothing but conditional beliefs held and not held by $\mathcal{A}$ in its initial state $[\rho, \blacktriangle]$. That is, o is a partial description of $\mathcal{A}$'s conditional beliefs in $[\rho, \blacktriangle]$. The proposition further entails that if we had a *full* description of its conditional beliefs we could calculate the beliefs after any sequence of revision inputs.

It turns out that the assumed belief revision framework allows us to apply existing work ([18] and in particular [8]) on completing partial information about conditional beliefs[2] and to construct a suitable ρ such that

[2] [8] presents a *rational closure* construction that takes into account both positive and negative information as is necessary in our case. It extends the case of positive-only information studied in [18]. These papers also inspired the name rational explanation.

$[\rho, \blacktriangle]$ is indeed an explanation for o in case $\blacktriangle$ is o-acceptable. $\rho_R(o, \blacktriangle)$ denotes the sequence thus constructed. The construction even reveals *if* a given core belief is o-acceptable.

We further showed that the set of o-acceptable cores is closed under disjunction. If $\blacktriangle_1$ and $\blacktriangle_2$ are o-acceptable, then so is $\blacktriangle_1 \vee \blacktriangle_2$.[3] This entails that—if o can be explained at all—there is a unique logically weakest o-acceptable core belief, which we denote by $\blacktriangle_\vee(o)$. Consequently $\blacktriangle \vdash \blacktriangle_\vee(o)$ for *any* o-acceptable $\blacktriangle$. The rationale behind choosing $\blacktriangle_\vee(o)$ for an explanation is that any input we predict to be rejected by $\mathcal{A}$ will indeed be rejected. Furthermore, it can be shown that adding beliefs or non-beliefs to o by strengthening some θ_i or enlarging some D_i as well as appending observations to the front or the back of o to get an observation o' cannot falsify this conclusion as $\blacktriangle_\vee(o') \vdash \blacktriangle_\vee(o)$. For any other core belief explaining o, a revision input predicted to be rejected by the agent might in fact be accepted. In this sense, we consider $\blacktriangle_\vee(o)$ to be optimal.

The choice of $\rho_R(o, \blacktriangle_\vee(o))$, which we call the *rational prefix*, as the sequence in the agent's initial epistemic state is justified by showing that it yields an optimal belief traces. Let $\rho = \rho_R(o, \blacktriangle_\vee(o))$ and σ be the sequence of any other explanation $[\sigma, \blacktriangle_\vee(o)]$ for o, $(\mathsf{Bel}_0^\rho, \mathsf{Bel}_1^\rho, \ldots, \mathsf{Bel}_n^\rho)$ and $(\mathsf{Bel}_0^\sigma, \mathsf{Bel}_1^\sigma, \ldots, \mathsf{Bel}_n^\sigma)$ be the corresponding belief traces. Then the following holds: If $\mathsf{Bel}_j^\rho \equiv \mathsf{Bel}_j^\sigma$ for all $j < i$ then $\mathsf{Bel}_i^\sigma \vdash \mathsf{Bel}_i^\rho$.[4] This tells us that the formulae we predict the agent to believe initially will indeed be believed (although some further formulae might be believed as well)—provided the agent's core belief really is $\blacktriangle_\vee(o)$. And if our predicted belief trace exactly captures the agent's beliefs up to the ith input then again all beliefs predicted after the next input will indeed be believed. The assumption that the two explanations use the same core belief causes this criterion, which we will refer to as the optimality criterion for the rational prefix, to be a rather weak one as we will see shortly.

In [7], we defined $[\rho_R(o, \blacktriangle_\vee(o)), \blacktriangle_\vee(o)]$ to be the *rational explanation* of an observation o—if there is an explanation at all. That paper and [5] contain more results about the rational explanation but these are the most important ones which justify the claim that the rational explanation is the best explanation for a given observation o. An algorithm which calculates the rational explanation is given below and described in more detail in [6]. The problem with calculating $[\rho_R(o, \blacktriangle_\vee(o)), \blacktriangle_\vee(o)]$ is that $\blacktriangle_\vee(o)$ has to be known, which it is not in the beginning. So the idea is to iteratively refine the core belief starting with the weakest possible of all $\top$.

[3] The proof is constructive and not deep but lengthy.

[4] This result, as almost all the others in this section, is proved in [5]. Note also that $\mathsf{Bel}_i^\rho \vdash \mathsf{Bel}_i^\sigma$ need not hold. Consider $[(\neg p), \top]$ and $[(p \wedge q, \neg p), \top]$. The belief traces when assuming a single input p are $(\neg p, p)$ and $(\neg p, p \wedge q)$. Although the beliefs are equivalent initially, they need not be after a revision step.

Algorithm 1: Calculation of the rational explanation.

Input: observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$
Output: the rational explanation for o

$\blacktriangle \Leftarrow \top$
repeat
  $\rho \Leftarrow \rho_R(o, \blacktriangle)$ /* now $\rho = (\alpha_m, \ldots, \alpha_0)$ */
  $\blacktriangle \Leftarrow \blacktriangle \wedge \alpha_m$
until $\alpha_m \equiv \top$
return $[\rho, \blacktriangle]$ if $\blacktriangle \not\equiv \bot$, "no explanation" otherwise

Having calculated the rational explanation $[\rho, \blacktriangle]$ of an observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$, we can make predictions concerning which inputs $\mathcal{A}$ accepts and rejects based on $\blacktriangle$ and our conclusions about its beliefs after having received each input are summarised by the corresponding belief trace $(\mathsf{Bel}_0^\rho, \mathsf{Bel}_1^\rho, \ldots, \mathsf{Bel}_n^\rho)$.

2.4 Safe conclusions and hypothetical reasoning

In the remainder of this section, we will illustrate some limitations of the rational explanation. In particular, we will show that predictions based on it will almost never be safe ones. However, this is inherent in the problem and not due to our solution to it.

As with many optimisation problems the quality of the solution and the conclusions we can draw from it depend heavily on the quality of the data and validity of the assumptions made. In our case, we clearly stated the assumptions made about the given observation as well as the agent's being ruled by the assumed framework. The optimality result for the best explaining core belief $\blacktriangle_\vee(o)$, i.e., that $\blacktriangle_\vee(o)$ is entailed by any o-acceptable core, depends on those. The optimality of the rational prefix $\rho_R(o, \blacktriangle)$ and therefore the conclusions about $\mathcal{A}'s$ further beliefs also depend on its actually employing the assumed core belief. That is, if we cannot be sure of the agent's actual core belief then most of what we can say about the agent's belief trace based on the rational explanation is merely justified guesses but not safe bets.

Example 2.9. *(i)* Let $o = \langle(\top, p, \varnothing), (\neg p, \top, \varnothing), (r \leftrightarrow \neg p, r \vee p, \varnothing)\rangle$. The rational explanation for o is $[(p), \top]$ and the corresponding belief trace is $(p, p, \neg p, r \wedge \neg p)$. That is, we conclude that $\mathcal{A}$ accepted the input $\neg p$ and believes $r \wedge \neg p$ after then receiving $r \leftrightarrow \neg p$.

Now assume the agent's real initial belief state was $[(), p]$—note that the core belief does not correspond to the one calculated by the rational explanation—and thus the belief trace in truth is $(p, p, p, \neg r \wedge p)$. That is, it did *not* accept the input $\neg p$ and believed $\neg r \wedge p$ after receiving $r \leftrightarrow \neg p$.

So except for the beliefs before the observation started and after receiving the tautology (where we are informed that the agent believes p and hence must have believed it initially) most of the conclusions about beliefs held by $\mathcal{A}$ we draw from the belief trace are wrong!

(ii) Let $o = \langle (p, p, \varnothing), (q, q, \varnothing), (r \leftrightarrow p, \top, \varnothing)\rangle$. The rational explanation for o is $[(), \top]$ and the belief trace implied by that explanation is $(\top, p, p \wedge q, p \wedge q \wedge r)$. Assuming that $[(), q \rightarrow \neg p]$ was $\mathcal{A}$'s true initial state, the belief trace in truth is $(q \rightarrow \neg p, p \wedge \neg q, q \wedge \neg p, q \wedge \neg p \wedge \neg r)$. Again, for large parts the conclusions we draw about the agent's beliefs based on the rational explanation are wrong. For example, we conclude that agent continues to believe p once it has been received. This is clearly not the case.

This strong dependence on the core belief can be easily explained. There are two main effects due to the core belief. First, it causes revision inputs to be rejected immediately. This is why the conclusions based on the rational explanation are off the mark in case *(i)* in the above example. Secondly, the core also accounts for interactions between revision inputs. An earlier input is eliminated from the belief set in the light of the core and some later inputs. This effect is illustrated in case *(ii)*. For one choice of the core belief, after having received the input φ_{i+j}, the agent may still believe the input φ_i received earlier, while for another core it may believe $\neg\varphi_i$.

Even if we got the core belief right and hence the agent really employs $\blacktriangle_\vee(o)$, conclusions based on the rational explanation of o should not be used without care. The optimality result for the rational prefix does not exclude mistakes. Correct conclusions about beliefs are guaranteed only up to the point in the belief trace where the beliefs we calculate and the agent's actual ones first fail to be equivalent. This can easily be the case already for the initial beliefs.

Consider $o = \langle (p, q, \varnothing), (r, \top, \varnothing)\rangle$ for which the rational explanation is $[(p \rightarrow q), \top]$, the corresponding belief trace being $(p \rightarrow q, p \wedge q, p \wedge q \wedge r)$. So we would conclude the agent to keep believing in q. If the agent's real initial epistemic state was $[(\neg q, \neg r \wedge q), \top]$ then the real belief trace would be $(\neg r \wedge q, p \wedge \neg r \wedge q, r \wedge \neg q)$. Although the correct core was calculated, we would still be wrong about q whose negation is in fact believed after having received the input r.

As stated above, using the rational explanation $[\rho, \blacktriangle_\vee(o)]$ we conclude that $\mathcal{A}$ believed $\mathsf{Bel}_i^\rho = f(\rho \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle_\vee(o)))$ after having received the first i revision inputs recorded in o. How safe is this conclusion? The above example showed that it is not very safe. So what can we do to further improve the results?

Here, we will consider only one very strong notion. We call the conclusion that $\mathcal{A}$ believes ψ after receiving the ith revision input recorded in o safe if and only if for *all* explanations for o we have $\mathsf{Bel}_i \vdash \psi$, where

Bel_i is the element of the belief trace corresponding to that input. In other words, every possible explanation predicts that belief (so in particular the one corresponding to the agent's real initial state). Analogously, we call the conclusion that the agent did *not* believe ψ at a certain point safe whenever *no* explanation predicts that formula to be believed. Note that a safe conclusion about an agent's belief does not mean that this belief is correct. The agent may have received and accepted unreliable information, but it means that given the observation, the agent must have held this belief.

We will now describe a way to calculate whether a conclusion of that form is safe, a method we call *hypothetical reasoning*. By this we mean modifying the given observation according to some conjecture and rerunning the rational explanation construction on the observation thus obtained. Note that any explanation for

$$o' = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_i, \theta_i \wedge \psi, D_i), \ldots, (\varphi_n, \theta_n, D_n)\rangle \text{ or}$$
$$o' = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_i, \theta_i, D_i \cup \{\psi\}), \ldots, (\varphi_n, \theta_n, D_n)\rangle$$

will also explain $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_i, \theta_i, D_i), \ldots, (\varphi_n, \theta_n, D_n)\rangle$. This follows directly from Definition 2.6. If $\theta_i \wedge \psi$ belongs to the beliefs after the ith revision step then so does θ_i and if none of the elements of $D_i \cup \{\psi\}$ is believed at that point, the same holds for any subset.

So in order to check whether the conclusion of $\mathcal{A}$ believing ψ after receiving the ith revision input is a safe one, we simply add ψ to D_i and test whether the observation thus obtained has an explanation.[5] If so, then the conclusion is not safe as there is an explanation where ψ is in fact not believed. However, if no such explanation exists then ψ must indeed be believed by $\mathcal{A}$. The non-belief of a formula ψ can be verified by replacing the corresponding θ_i by $\theta_i \wedge \psi$. If the observation thus obtained has an explanation then the agent may have believed ψ and consequently the conclusion is not safe.

With a small modification this method works also for hypothetical reasoning about the agent's initial beliefs, i.e., before receiving the first input. It does not work directly, as the observation does not contain an entry for the initial state. By appending $\langle(\varphi_0, \theta_0, D_0)\rangle = \langle(\top, \top, \varnothing)\rangle$ to the front of the observation o we create this entry. The point is that receiving a tautology as input leaves the beliefs unchanged. We can now add a formula ψ to D_0 or θ_0 as described above to verify conclusions about the initial state. Further, there is no restriction which formulae ψ can be used for hypothetical reasoning. It is even possible to add several ψ_j simultaneously to o to get a modified observation o'.

[5] The rational explanation algorithm always finds an explanation if there is one, and returns "no explanation" if there is none.

Hypothetical reasoning can also be used in order to improve the conclusions about $\mathcal{A}$'s core belief $\blacktriangle$. We already know that $\blacktriangle \vdash \blacktriangle_\vee(o)$, i.e., all inputs we predict to be rejected by $\mathcal{A}$ will indeed be rejected. This is because any o-acceptable core entails $\blacktriangle_\vee(o)$. But what about the other inputs, must they really have been accepted? Can we be sure that φ_i really was accepted if it is consistent with $\blacktriangle_\vee(o)$? Rejecting φ_i is equivalent to not believing the input after having received it. So, we simply add φ_i to D_i, i.e., replace $(\varphi_i, \theta_i, D_i)$ in o by $(\varphi_i, \theta_i, D_i \cup \{\varphi_i\})$ to get o'. If there is an o'-acceptable core, then $\mathcal{A}$ may in fact have rejected φ_i. However, if o' does not have an explanation then we know that $\mathcal{A}$ must have accepted that input.

It might be nice to be able to check whether conclusions about $\mathcal{A}$'s beliefs are safe, but can we ever be sure to have the correct core belief in order to apply the optimality results we gave for the rational prefix? The answer to this question is almost exclusively negative. Usually, there is more than one o-acceptable core. In a different context Sébastien Konieczny[6] suggested the additional assumption that the last belief θ_n recorded in the observation o is in fact *complete*. This assumption gives us an upper bound on the actual core belief $\blacktriangle$ as then $\theta_n \vdash \blacktriangle \vdash \blacktriangle_\vee(o)$ must hold and we can use the hypothetical reasoning methodology in order to get an improved core belief. As we know the exact belief θ_n at the end of the observation, we can iteratively add to D_n those formulae ψ which the rational explanation predicts to be believed but which are not entailed by θ_n. This method will yield an improved lower bound for the core belief of the agent, but it cannot guarantee uniqueness of the core.

Even if we assumed that *every* θ_i completely characterises the beliefs of the agent after receiving φ_i, we would not be guaranteed to get the real core belief. Consider $o = \langle (p, p, \varnothing), (q, p \wedge q, \varnothing), (r, p \wedge q \wedge r, \varnothing) \rangle$ to illustrate this. The rational explanation for o is $[(), \top]$. However, p is also an o-acceptable core, $[(), p]$ being one possible explanation. That is, the conclusion that an input $\neg p$ will be accepted by the agent is not safe. This illustrates that even using much more severe assumptions about a given observation, identifying the agent's real core belief is impossible.

3 Extension to Unknown Subformulae

Up to this point we considered observations that were complete with respect to the revision inputs received. We knew exactly which inputs were received during the time of observation. The scenarios in the introduction suggested that it is well possible that some of the inputs might have been missed. Further, the observer may not understand the complete *logical content* of all the revision inputs, $\mathcal{A}$'s beliefs and non-beliefs. Consider the following example

[6] Personal communication.

where the agent is observed to receive *exactly* two inputs p and q. After hearing p, the agent believed *something we cannot understand*, but after then hearing q, it did not believe that anymore. In the original framework this cannot be formalised as there is no means to represent the unknown belief. However, we should be able to conclude that $\mathcal{A}$ believed $\neg q$ after having received p. This is because the assumed belief revision framework satisfies (most of) the AGM postulates [1]. In particular, if the input is consistent with the current beliefs they have to survive the revision process (cf. the "Vacuity" postulate from AGM) which is clearly not the case in the example. The current section investigates how the previous results can still be used to reason about $\mathcal{A}$ if the observations are allowed to be less complete in this sense.

We want to emphasise that there is a big difference between knowing there was a revision input while being ignorant about its logical content and not even knowing whether there were one or more revision inputs. We will first deal with the former case which will provide results to deal with the latter one in Section 3.3.

3.1 Modelling unknown logical content

We will model partial information by allowing formulae appearing in the observation to contain unknown subformulae which are represented by n placeholders χ_j. $\lambda(\chi_1, \ldots, \chi_n)[(\chi_i/\phi_i)_i]$ denotes the result of replacing in λ every occurrence of χ_i by ϕ_i.

Definition 3.1. Let L be a propositional language and $\chi_1, \ldots, \chi_n$ be placeholders not belonging to L.

A "formula" $\lambda(\chi_1, \ldots, \chi_n)$ possibly containing $\chi_1, \ldots, \chi_n$ is called a *parametrised formula based on* L iff $\lambda(\chi_1, \ldots, \chi_n)[(\chi_i/\phi)_i] \in L$ whenever $\phi \in L$. $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_l, \theta_l, D_l)\rangle$ is a *parametrised observation based on* L iff all φ_i, θ_i, $\delta \in D_i$ are parametrised formulae based on L. We denote by $L(o)$ the smallest language L a parametrised observation o is based on.

To put it differently, a parametrised formula based on L is a formula from L in which some subformulae have been replaced by placeholders χ_i. This allows hiding parts of the logical content of a formula. So in order to model (even more) partial knowledge, we will consider parametrised observations. The example from the introductory paragraph can now be represented by $o = \langle(p, \chi, \varnothing), (q, \top, \{\chi\})\rangle$. We will often write λ rather than $\lambda(\chi_1, \ldots, \chi_n)$ to denote a parametrised formula in order to ease reading.

Unknown subformulae χ_i are allowed to appear in all components of an observation—revision inputs, beliefs and non-beliefs. The same χ_i can appear several times. In fact, this is when it contributes to the reasoning

process. It is not unreasonable to assume that this can happen. For example, the meaning of an utterance in a dialogue might not be understood as part of the language may not be known to the observing agent, but the utterance might be recognised when it appears again later. Analogous to a learner of a foreign language, the observer may be familiar with (parts of) the structure of the language while being ignorant about the meaning of certain "phrases". In case we are completely ignorant about the logical content the entire parametrised formula will simply be a placeholder.

Let o be a parametrised observation. $o[\chi_1/\phi_1, \ldots, \chi_n/\phi_n]$ and equivalently $o[(\chi_i/\phi_i)_i]$ denote the observation obtained by replacing in o every occurrence of the placeholder χ_i by a formula ϕ_i.

We still assume correctness of the information contained in the parametrised observation o, i.e., we assume the existence of instantiations ϕ_i of all unknown subformulae χ_i such that the observation $o[(\chi_i/\phi_i)_i]$ is a correct observation in the sense of Section 2—in particular, there must be an entry for every revision input received. The agent indeed received exactly the inputs recorded and beliefs and non-beliefs are correct if partial. Note that this implies that we are not yet able to deal with *missing* inputs. These will be considered in Section 3.3. One important technical restriction is that the instantiations of unknown subformulae χ_i must not contain unknown subformulae χ_j themselves, i.e., the instantiations must be elements of the underlying language—however, not necessarily elements of $L(o)$. That is, the true meaning of χ_i is not assumed to be expressible in the language of the known part of o. Abusing notation we will write that o has an explanation, meaning that there exist instantiations $\phi_1, \ldots, \phi_n$ for the unknown subformulae such that $o[(\chi_i/\phi_i)_i]$ has an explanation; similarly that $\blacktriangle$ is o-acceptable if $\blacktriangle$ is $o[(\chi_i/\phi_i)_i]$-acceptable.

3.2 Finding an acceptable core belief

In this section, we will present results on what can be said about $\mathcal{A}$'s core belief given a parametrised observation o. If an explanation exists at all, once more there will be a unique weakest o-acceptable core $\blacktriangle$. This may be surprising as there are many different possible instantiations for the unknown subformulae. But this will also allow us to choose them such that any o-acceptable core entails $\blacktriangle$. If we knew the instantiations of the unknown subformulae we could simply use the rational explanation algorithm, as in that case a parametrised observation could be transformed into a regular one. As we do not know them, we have to guess. The trick is to extend the language and treat every χ_i as a new propositional variable x_i.

Proposition 3.2. If $[\rho, \blacktriangle]$ explains $o[(\chi_i/\phi_i)_i]$ and $x_1, \ldots, x_n$ are propositional variables not appearing in o, $\blacktriangle$, ρ or any ϕ_i then $o[(\chi_i/x_i)_i]$ is explained by $[\rho, \blacktriangle \wedge \bigwedge_{1 \leq i \leq n}(x_i \leftrightarrow \phi_i)]$.

Proof (Sketch). $\lambda[(\chi_i/\phi_i)_i] \vdash \bot$ iff $\bigwedge_{1\leq i\leq n}(x_i \leftrightarrow \phi_i) \wedge \lambda[(\chi_i/x_i)_i] \vdash \bot$ for any parametrised formula λ not containing x_i is the key to this result. As the x_i are not contained in $[\rho, \blacktriangle]$ or $o[(\chi_i/\phi_i)_i]$, requiring $\bigwedge(x_i \leftrightarrow \phi_i)$ ensures that the different instantiations have the same logical consequences—modulo entailment of irrelevant formulae containing the x_i. The (relevant) beliefs are the same for both explanations. Q.E.D.

The proposition formalises that given there is *some* instantiation for the unknown subformulae in o such that the resulting observation has an explanation, we can also replace them by new variables and still know that there is an explanation. However, this tells us that we can apply the rational explanation algorithm to $o[(\chi_i/x_i)_i]$ and be guaranteed to be returned an explanation if there is one. If this fails, i.e., we are returned an inconsistent core belief, then no explanation can exist using any instantiation of the unknown subformulae in o. The core belief $\blacktriangle \wedge \bigwedge_{1\leq i\leq n}(x_i \leftrightarrow \phi_i)$ ensures that the new variables x_i behave exactly as the "correct" instantiations ϕ_i for the unknown subformulae in o.

In general, $\blacktriangle_\vee(o[(\chi_i/x_i)_i])$—the core belief returned by the rational explanation algorithm—will not be that particular formula. Note that this would be impossible as there can be several suitable instantiations such that $o[(\chi_i/\phi_i)_i]$ has an explanation. The core belief calculated will in general be weaker but may still contain (some of) the additional variables x_i. We will now go on to show that it is possible to eliminate these variables from the core belief by choosing different instantiations for the unknown subformulae.

The idea is to split the core $\blacktriangle$ calculated by the rational explanation construction into two parts, one $\blacktriangle'$ that talks only about $L(o)$ and not at all about the additional variables and one ψ part that talks also about those. Formally, we choose $\blacktriangle'$ and ψ such that $\blacktriangle \equiv \blacktriangle' \wedge \psi$ and $\mathsf{Cn}(\blacktriangle') = \mathsf{Cn}(\blacktriangle) \cap L(o)$, which is possible as we are in a finite setting.[7] Instead of x_i we then use $x_i \wedge \psi$ to instantiate the placeholders. This shifting of parts of the core to the new variables is possible because the part of the core belief that talks about the x_i becomes relevant in the calculation of the beliefs of an agent only when those variables themselves appear.

Proposition 3.3. If $[\rho, \blacktriangle]$ explains $o[(\chi_i/x_i)_i]$ then there exist $\blacktriangle'$ and ψ such that $\blacktriangle'$ contains no x_i and $[\rho \cdot \psi, \blacktriangle']$ explains $o[(\chi_i/x_i \wedge \psi)_i]$.

Proof (Sketch). Let $\psi = \blacktriangle$ and $\blacktriangle'$ such that $\mathsf{Cn}(\blacktriangle') = \mathsf{Cn}(\blacktriangle) \cap L(o)$. It can now be shown that $f(\varphi_1[(\chi_i/x_i)_i], \ldots, \varphi_j[(\chi_i/x_i)_i], \blacktriangle)$ is equivalent to $f(\blacktriangle, \varphi_1[(\chi_i/x_i \wedge \blacktriangle)_i], \ldots, \varphi_j[(\chi_i/x_i \wedge \blacktriangle)_i], \blacktriangle')$. Again, the proof of that is

[7] We can trivially choose $\psi = \blacktriangle$ and $\blacktriangle'$ to represent all logical consequences of $\blacktriangle$ in $L(o)$.

not deep but lengthy. The intuition is that $\blacktriangle'$ makes sure that with respect to $L(o)$ all formulae are treated correctly and using $x_i \wedge \blacktriangle$ rather than just x_i, the effect of $\blacktriangle$ with respect to new variables is maintained. Consequently, before processing ρ when calculating the beliefs, in both cases equivalent formulae have been constructed and the beliefs will thus be equivalent. Q.E.D.

To summarise what we know so far. Given a parametrised observation o has an explanation, we can construct one for $o[(\chi_i/x_i)_i]$. However, the corresponding core belief $\blacktriangle_\vee(o[(\chi_i/x_i)_i])$ may still contain variables that are not contained in $L(o)$ and thus we cannot claim that the agent had contact with them. The last proposition now showed that we can construct instantiations for the unknown subformulae such that the explaining core belief $\blacktriangle'$ is in $L(o)$. We can even go one step further and show that *any* $o[(\chi_i/\phi_i)_i]$-acceptable core $\blacktriangle''$ must entail the core $\blacktriangle'$ constructed as described above.

Proposition 3.4. Let $[\rho'', \blacktriangle'']$ be an explanation for $o[(\chi_i/\phi_i)_i]$ and $[\rho, \blacktriangle]$ be the rational explanation for $o[(\chi_i/x_i)_i]$, where x_i are additional propositional variables not appearing in any ϕ_i, $\blacktriangle''$ or the language $L = L(o)$. Further let $\blacktriangle'$ such that $\mathsf{Cn}(\blacktriangle') = \mathsf{Cn}(\blacktriangle) \cap L$. Then $\blacktriangle'' \vdash \blacktriangle'$.

Proof. By Proposition 3.2 $\blacktriangle'' \wedge \bigwedge(x_i \leftrightarrow \phi_i)$ is $o[(\chi_i/x_i)_i]$-acceptable and hence entails $\blacktriangle$ (any o-acceptable core entails $\blacktriangle_\vee(o)$). Obviously $\blacktriangle \vdash \blacktriangle'$, so $\blacktriangle'' \wedge \bigwedge(x_i \leftrightarrow \phi_i) \vdash \blacktriangle'$. Now assume $\blacktriangle''$ does not entail $\blacktriangle'$ which implies there is a model for $\blacktriangle'' \wedge \neg\blacktriangle'$. Neither $\blacktriangle''$ nor $\blacktriangle'$ contain any x_i so we can extend that model to one for $\blacktriangle'' \wedge \bigwedge(x_i \leftrightarrow \phi_i) \wedge \neg\blacktriangle'$ by evaluating x_i just as ϕ_i—contradicting $\blacktriangle'' \wedge \bigwedge(x_i \leftrightarrow \phi_i) \vdash \blacktriangle'$. Q.E.D.

There is an important consequence of that result. As in the original case there is a unique weakest o-acceptable core for a parametrised observation o. This follows directly from the last two propositions. $\blacktriangle'$, being constructed as described above, is o-acceptable and is entailed by any o-acceptable core, so in particular by the agent's real core belief. Hence, all formulae inconsistent with $\blacktriangle'$ will be rejected by $\mathcal{A}$. That is, $\blacktriangle'$ yields a safe conclusion with respect to which formulae must be rejected by $\mathcal{A}$—no matter what the instantiations of the unknown subformulae really are.

Example 3.5. Consider $o = \langle(\chi, \chi, \varnothing), (p, q \wedge \neg\chi, \varnothing)\rangle$. This parametrised observation expresses that the observed agent accepted an input whose meaning is unknown to us. After then receiving p, it believed q and the negation of the unknown input. The observation constructed according to Proposition 3.2, where χ is replaced by a new variable x, is $o[\chi/x] = \langle(x, x, \varnothing), (p, q \wedge \neg x, \varnothing)\rangle$. The rational explanation for $o[\chi/x]$ is

$[(p \wedge \neg x \rightarrow q), p \rightarrow \neg x]$ and $(p \rightarrow (\neg x \wedge q), x \wedge \neg p, p \wedge q \wedge \neg x)$ is the corresponding belief trace.

This indicates that after receiving the unknown input the agent believes $\neg p$. In order to test whether this is necessarily the case, we investigate the parametrised observation $o' = \langle(\chi, \chi, \{\neg p\}), (p, q \wedge \neg\chi, \varnothing)\rangle$. According to the hypothetical reasoning methodology, $\neg p$ was added to the non-beliefs. Applying the rational explanation algorithm yields that $o'[\chi/x]$ has no explanation. Proposition 3.2 now tells us that there cannot be an explanation for o'—no matter how χ is instantiated. That is, if the parametrised observation correctly captures the information about the agent, it must believe $\neg p$ after receiving the first input.

o is based on the language L constructed from the variables p and q and $\mathsf{Cn}(p \rightarrow \neg x) \cap L = \mathsf{Cn}(\top)$. To illustrate Propositions 3.3 and 3.4 note that $o[\chi/x \wedge (p \rightarrow \neg x)] = \langle(x \wedge \neg p, x \wedge \neg p, \varnothing), (p, q \wedge (\neg x \vee p), \varnothing)\rangle$ is explained by $[(p \wedge \neg x \rightarrow q, p \rightarrow \neg x), \top]$, the corresponding belief trace being $(p \rightarrow (\neg x \wedge q), x \wedge \neg p, p \wedge q \wedge \neg x)$. $\top$ is trivially entailed by any o-acceptable core.

In order to find an acceptable core for a parametrised observation o, we extended the language $L(o)$ with new variables. In [23], we gave an example—which we will not repeat here—illustrating that there are parametrised observations that have an explanation when language extension is allowed but which cannot be explained restricting the language to $L(o)$. In other words, the proposed algorithm of replacing each χ_i by a new variable x_i, running the rational explanation construction and then eliminating the x_i from the core belief (the result being $\blacktriangle'$) may yield an explanation, although none exists when restricting the instantiations of the χ_i to $L(o)$. Although we know that each acceptable core will entail $\blacktriangle'$, we cannot generally say that restricting the instantiations to $L(o)$ will allow the same core, a strictly stronger one or none at all to explain o.

Note that Proposition 3.2 makes no assumption about the language of the instantiations ϕ_i of the unknown subformulae. They may or may not belong to $L(o)$. They may contain arbitrarily (but finitely) many propositional variables not belonging to $L(o)$. However, that proposition has an interesting implication. It says if $o[(\chi_i/\phi_i)_i]$ has an explanation then so does $o[(\chi_i/x_i)_i]$, but $o[(\chi_i/x_i)_i]$ contains only variables from $L(o)$ and n additional variables x_i, one for each placeholder. As that observation has an explanation, the rational explanation construction will return one. However, that construction uses only formulae present in the observation. Consequently, it does not invent new variables. So, no matter how many variables not appearing in $L(o)$ were contained in the ϕ_i, n additional variables suffice for finding an explanation for the parametrised observation o. This yields an upper bound on additional variables needed.

In Section 2.4 we showed that assuming the wrong core belief greatly affects the quality of the conclusions about $\mathcal{A}$'s other beliefs. And even if the core is correct, the belief trace implied by the rational explanation does not necessarily yield only safe conclusions with respect to the beliefs of the agent during the observation.

These problems are obviously inherited by the current extension to partial information about the logical content of the formulae in an observation. They cannot be expected to become less when not even knowing what inputs the agent really received or when information about the beliefs and non-beliefs becomes even more vague. Much depends not only on the core belief but also on the instantiation of the unknown subformulae. So rather than just having to calculate a best initial epistemic state, we now would also have to find an optimal instantiation of the unknown subformulae. However, the limitations illustrated in Section 2.4 prevent us from even attempting to look for them. Instead, we propose to investigate the belief trace implied by the rational explanation of an observation $o[(\chi_i/x_i)_i]$ and reason hypothetically about beliefs and non-beliefs from $L(o)$ in that belief trace.

3.3 Intermediate inputs

Up to now, we assumed the (parametrised) observation o to contain an entry (φ, θ, D) for every revision input received by $\mathcal{A}$, even if some of the formulae are only partially known. This corresponds to the assumption of having an eye on the agent at all times during the observation. In this section, we want to drop this assumption. That is, we will allow for intermediate inputs between those recorded in o. In real applications this will be the norm rather than an exceptional case. $\mathcal{A}$ or the observing agent may leave the scene for a time, and if the observing agent is the source of information then o might have been gathered over several sessions between which $\mathcal{A}$ may have received further input.

Using our notation for observations, an intermediate input is one we have no information about, i.e., we do not know what the revision input is or what is believed or not believed after receiving it. Hence, we can represent it by $\langle(\chi, \top, \varnothing)\rangle$; χ again represents an unknown formula. Note that this is different from $\langle(\chi, \chi, \varnothing)\rangle$ as here the input would be required to be accepted by $\mathcal{A}$. In other words, the agent's core belief would have to be consistent with the instantiation of χ.

Example 3.6. Consider the following observation without intermediate inputs: $o = \langle(p, q, \varnothing), (p, \neg q, \varnothing)\rangle$. Assume $\blacktriangle$ was o-acceptable and thus consistent. Then either it is consistent or inconsistent with p. In both cases, the belief set does not change upon receiving the second input p. Either the first p was accepted and hence already believed or p was rejected (both

times) in which case the belief set never changes. So $\neg q$ must have been believed already after the first p was received. But it is not possible to believe $q \wedge \neg q$ consistently (the belief set is inconsistent if and only if the core belief is inconsistent). Consequently, there is no o-acceptable core.

Assuming a single intermediate input $\langle(\chi, \top, \varnothing)\rangle$, there is only one reasonable position yielding $o' = \langle(p, q, \varnothing), (\chi, \top, \varnothing), (p, \neg q, \varnothing)\rangle$. Instantiating the unknown formula χ with $p \rightarrow \neg q$, $[(p \rightarrow q), \top]$ is an explanation. Before receiving the first input the agent believes $p \rightarrow q$, after receiving the first p it believes $p \wedge q$ and after receiving the (assumed) intermediate input as well as after receiving the last input p it believes $p \wedge \neg q$. Hence $\top$ is o'-acceptable. That is, while o does not have an explanation, assuming an intermediate input allows the observation to be explained.

In the general case we do not know how many intermediate inputs were received at which points in a given (parametrised) observation o. In [23] we showed that number and positions of the intermediate inputs have an impact on the possible explanations of o. If number and positions are fixed then we deal with a parametrised observation (containing an entry for every revision input received) and can hence use the results of Section 3.2 in order to calculate the weakest acceptable core belief. To represent the intermediate inputs we simply have to introduce *further* unknown subformulae not contained in o. Assume we have the partial observation $o = \langle(p, q \wedge \chi_1, \varnothing), (r, \neg q, \varnothing), (p, q, \{\chi_1\})\rangle$ and the information that exactly two intermediate inputs have been received immediately after r. In order to reason about $\mathcal{A}$, we consider the partial observation $o' = \langle(p, q \wedge \chi_1, \varnothing), (r, \neg q, \varnothing), (\chi_2, \top, \varnothing), (\chi_3, \top, \varnothing), (p, q, \{\chi_1\})\rangle$ which now contains an entry for every input received. At this point we want to emphasise once more that intermediate inputs and partial information about inputs are related but distinct cases.

In the following we want to indicate what can be said about the agent's core belief depending on how much information we have concerning possible intermediate inputs. Naturally, the more specific our knowledge concerning number and positions, the more informative the conclusions can be. We will start with the case where we have no information at all, which means that any number of intermediate inputs may have been received any time. Then we will turn to the cases where the positions or the number are restricted.

Any number of intermediate inputs at any time. Consider an observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_n, \theta_n, D_n)\rangle$. Assume $[\rho, \blacktriangle]$ explains the observation o' which is obtained from o by putting some arbitrary number of intermediate inputs at any position in o. It can be proved that then there are a sequence σ and $n-1$ intermediate inputs such that $[\sigma, \blacktriangle]$ explains o'' obtained from o by putting exactly one intermediate input between any two inputs φ_i and φ_{i+1} in o. Note that both explanations use the same core

belief. Intuitively, the intermediate input from o'' before the input φ_{i+1} is the conjunction of *all relevant* intermediate inputs from o' before that input.

Proposition 3.2 tells us that we can also use new variables x_i instead of those intermediate inputs[8] and the observation o''' thus obtained is guaranteed to have an explanation. However, o''' does not contain any unknown subformulae, so we can apply the rational explanation construction which will return some epistemic state with core belief $\blacktriangle'$. We can now construct the weakest possible core belief by taking $\blacktriangle''$ such that $\mathsf{Cn}(\blacktriangle'') = \mathsf{Cn}(\blacktriangle') \cap L(o)$. Any o'-acceptable core belief—o' being constructed as described above—will entail $\blacktriangle''$. That is from $\blacktriangle''$ we can safely conclude which formulae are rejected by $\mathcal{A}$, no matter how many intermediate inputs it received at any point during the observation.

What happens if we have further information about the positions or the number of intermediate inputs? The following proposition implies that we should always assume the maximal number of intermediate inputs. It says that an additional intermediate input, which we instantiate with a new variable for calculating the weakest possible core belief, can only make the core logically weaker. Conversely, not assuming the maximal number of intermediate inputs may lead to the conclusion that $\mathcal{A}$ rejects a formula which it actually does not reject simply because an additional intermediate input allows $\mathcal{A}$'s core belief to be logically weaker.

Proposition 3.7. If $\mathsf{Cn}(\blacktriangle) = \mathsf{Cn}(\blacktriangle_\vee(o_1 \cdot \langle(x, \top, \varnothing)\rangle \cdot o_2)) \cap L(o_1 \cdot o_2)$ and $x \notin L(o_1 \cdot o_2)$ then $\blacktriangle_\vee(o_1 \cdot o_2) \vdash \blacktriangle$.

Proof (Sketch). By showing $\blacktriangle_\vee(o_1 \cdot o_2) \equiv \blacktriangle_\vee(o_1 \cdot \langle(\top, \top, \varnothing)\rangle \cdot o_2)$, which holds because a tautologous input has no impact, we introduce the extra input which allows us to compare the cores. By Proposition 3.2, $\blacktriangle_\vee(o_1 \cdot \langle(\top, \top, \varnothing)\rangle \cdot o_2) \wedge x$ is $o_1 \cdot \langle(x, \top, \varnothing)\rangle \cdot o_2$-acceptable and hence entails $\blacktriangle_\vee(o_1 \cdot \langle(x, \top, \varnothing)\rangle \cdot o_2)$. We can now show that any formula from $L(o_1 \cdot o_2)$ entailed by that core as already entailed by $\blacktriangle_\vee(o_1 \cdot o_2)$. Q.E.D.

Fixed positions of intermediate inputs. Now assume we know the positions where intermediate inputs may have occurred. This is imaginable, for example, in scenarios where the observing agent gathers o in several sessions, but does not know if $\mathcal{A}$ receives further inputs between those sessions. How many intermediate inputs should be assumed at each of those points? We cannot allow an arbitrary number as this is computationally infeasible, so it would be helpful to have an upper bound which we could then use. We claim that it suffices to assume j intermediate inputs at a particular position in o, where j is the number of revision inputs recorded in o following

[8] That is, we put an entry $\langle(x_i, \top, \varnothing)\rangle$ with a new variable x_i between any two entries $\langle(\varphi_i, \theta_i, D_i)\rangle$ and $\langle(\varphi_{i+1}, \theta_{i+1}, D_{i+1})\rangle$ in o.

that position, i.e., ignoring possible intermediate inputs appearing later.[9] The intuition is as above. For every recorded revision input, we assume one intermediate input which collects all the relevant intermediate inputs that have really occurred.

If this claim is correct, we can introduce into o one entry $(\chi_i, \top, \varnothing)$ for every intermediate input. Thus we get a parametrised observation containing an entry for every revision input received. We can then construct a weakest acceptable core belief by instantiating each χ_i by x_i, calculating the rational explanation of the observation thus obtained and then eliminating the additional variables from the core belief. For example, given an observation $o = \langle(\varphi_1, \theta_1, D_1), \ldots, (\varphi_5, \theta_5, D_5)\rangle$ and the information that intermediate inputs have been received only after φ_2 and φ_4, we can calculate the weakest possible core starting with $o' = \langle(\varphi_1, \theta_1, D_1), (\varphi_2, \theta_2, D_2), (x_1, \top, \varnothing), (x_2, \top, \varnothing), (x_3, \top, \varnothing), (\varphi_3, \theta_3, D_3), (\varphi_4, \theta_4, D_4), (x_4, \top, \varnothing), (\varphi_5, \theta_5, D_5)\rangle$ and eliminating the x_i from $\blacktriangle_\vee(o')$. Again, all x_i are propositional variables not contained in $L(o)$.

The above claim for limiting the number of assumed intermediate inputs follows almost immediately from the following proposition.

Proposition 3.8. Let $\rho = (\varphi_1, \ldots, \varphi_n)$ and $\sigma = (\psi_1, \ldots, \psi_m)$. Then there exists a $\sigma' = (\psi'_1, \ldots, \psi'_n)$ such that for all $1 \leq i \leq n$

$$f(\sigma \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle)) \equiv f(\sigma' \cdot (\varphi_1, \ldots, \varphi_i, \blacktriangle)) .$$

Proof (Sketch). The proof of this result uses the fact that for every sequence σ there is a logical chain σ' (a sequence of formulae where each formula is entailed by its successor) that behaves exactly like σ. That is $f(\sigma \cdot \rho') \equiv f(\sigma' \cdot \rho')$ for all sequences ρ'. However, for this result it suffices that σ and σ' behave equivalently for all prefixes of ρ. We then show that a suitable σ' exists, in fact using the rational explanation algorithm and hypothetical reasoning. Q.E.D.

Note that this result is not trivial, as m can be (much) greater than n and in this case we have to find a shorter sequence yielding equivalent formulae for all $1 \leq i \leq n$. This proposition tells us that we can replace one block of intermediate inputs σ by one of the proposed length and be guaranteed an equivalent formula being constructed in the calculation for each recorded revision input φ_i coming later in the observation.

We want to remark that some care has to be taken when considering the general case, where several blocks of intermediate inputs exist. Then ρ in the proposition may contain more elements than just the recorded revision

[9] The above result—that one intermediate input between any two recorded ones is enough—is not applicable here. Intermediate inputs may not be allowed at every position.

inputs; it also contains intermediate ones. And thus we have to find a sequence σ' not of length n but $j \leq n$ where j is the number of recorded inputs. We are currently investigating whether the number of intermediate inputs that have to be assumed can be reduced further without effect on the core belief calculated.

Fixed number of intermediate inputs. If we are given a maximal (or exact) number n of intermediate inputs that may have occurred we can draw conclusions about the core belief of the agent using the following method. Due to Proposition 3.7 we should indeed assume the maximal number of intermediate inputs—n. So let o be the observation containing only recorded inputs. If o has less than $n+2$ recorded inputs and there are no restrictions as to the positions of the intermediate inputs, we can use the result that one intermediate input between any two recorded ones suffices to explain o; otherwise, there are not enough intermediate inputs for this result to be applicable. In this case, we create the set of all possible observations o' where n intermediate inputs have been inserted in o:

$$O' = \{o_1 \cdot \langle (x_1, \top, \varnothing) \rangle \cdot o_2 \cdot \ldots \cdot \langle (x_n, \top, \varnothing) \rangle \cdot o_{n+1} \mid o = o_1 \cdot \ldots \cdot o_{n+1}\}.$$

Here we have already replaced the unknown formulae by new variables. If we have information about the positions of the intermediate inputs we can also take this into account when constructing O'. The observation o_j may be empty, so consecutive intermediate inputs are explicitly allowed. Now any possible core belief will entail $\bigvee\{\blacktriangle \mid \mathsf{Cn}(\blacktriangle) = \mathsf{Cn}(\blacktriangle_\vee(o')) \cap L(o), o' \in O'\}$. Note that this formula itself need not be an o'-acceptable core, i.e., it may not really explain the observation using n intermediate inputs. Conclusions about beliefs and non-beliefs can only be safe if they are safe for every observation in O'.

3.4 Summary

In this section, we showed what can still be said about $\mathcal{A}$ if some of the completeness assumptions about the observation o are weakened. We started by allowing unknown subformulae χ_i to appear in o. This can happen as the logical content of the revision inputs or the beliefs need not be completely known. In case the observation still contains a record for every input received, the calculation of an optimal core belief is still possible. The proposed method for dealing with such parametrised observations was to instantiate the unknown subformulae χ_i with new variables and apply the rational explanation construction to the observation thus obtained. From this explanation we can safely conclude which beliefs must belong to the agents core belief no matter what the real instantiation of the χ_i was.

We showed in [23] that although we can construct a core belief from $L(o)$ this does not guarantee that o can be explained without extending the

language. The unknown subformulae may still have to contain variables not belonging to $L(o)$. We claim that it is not useful to look for an optimal instantiation of the unknown subformulae. Weakest core belief and belief trace heavily depend on the choice of the instantiation of the χ_i and even if we had the correct ones, Section 2.4 showed that the conclusions drawn from the belief trace implied by our explanation are of limited use. Instead we argue that the χ_i should be instantiated with x_i and reasoning be done based on the rational explanation. This allows us to draw correct conclusions about the actual core belief of the agent, which must entail the one calculated that way. Further, we can use hypothetical reasoning to verify other beliefs and non-beliefs (restricted to $L(o)$) implied by the explanation thus obtained.

The additional assumption that the belief corresponding to the last revision input in the (parametrised) observation completely characterises $\mathcal{A}$'s beliefs at that point once more need not help. It might not even convey additional information about the language of the agent's epistemic state or of the unknown subformulae. Consider the parametrised observation $\langle(p \wedge \chi, \top, \varnothing), (\neg p, \neg p, \varnothing)\rangle$. It might not be very interesting but it illustrates the point. As $\neg p$ is inconsistent with the first input, χ could be instantiated with any formula and still $\neg p$ would completely characterise the agent's final beliefs.

We then further allowed intermediate inputs, i.e., the original observation does not contain a record for every input received. Some observations can be explained only when assuming that intermediate inputs have occurred. When fixing their number and positions, the problem is reduced to partially known inputs. If the observing agent does not have this information, we sketched procedures for drawing conclusions about what $\mathcal{A}$'s core belief must entail.

4 Conclusion, Future and Related Work

In this paper, we departed from the traditional belief revision setting of investigating what an agent should believe after receiving (a sequence of pieces of) new information in a given initial state. Instead, we place ourselves in the position of an observer trying to reason about another agent in the process of revising its beliefs. Coming up with models of other agents is useful in many application areas as informed decisions may improve the personal or group outcome of interactions.

The basic and admittedly oversimplified setting we consider is that we are given an observation containing propositional information about the revision inputs received by an agent $\mathcal{A}$ and about its beliefs and non-beliefs following each input. We investigated several degrees of incompleteness of the information provided. From such an observation we try to get a clearer

picture of $\mathcal{A}$. Assuming $\mathcal{A}$ to employ a particular belief revision framework, the general approach for reasoning about the agent is to "regress" the information contained in the observation to arrive at a possible initial state of the agent. This state completely determines the revision behaviour and therefore allows to draw conclusions about $\mathcal{A}$'s beliefs at each point in time during the observation as well as future beliefs. Even under the very strict assumptions we impose, hardly any safe conclusions can be drawn. Intuitively, this is because coming up with $\mathcal{A}$'s true initial state is virtually impossible. The observing agent can only try to extend and refine the observation and reason hypothetically in the sense of testing conjectures about $\mathcal{A}$ in order to improve the model.

It should be clear that the general question does not require the use of the belief revision framework we assumed. For future work, it might be interesting to see if similar results can be obtained when assuming $\mathcal{A}$ to employ a different framework. It would be interesting to see how different revision frameworks compare with respect to their power to explain an observation and whether there is a significant difference in the quality of the conclusions that can be drawn. Another important question is whether there is a way to actually find out which revision framework an observed agent employs or whether other assumptions can be verified. We claimed that it is not reasonable to look for the optimal instantiation of the unknown subformulae but rather do hypothetical reasoning restricted to $L(o)$. However, in some applications it might be interesting to know what the actual revision input was that triggered a certain reaction in the agent. So comparing potential instantiations (possibly from a fixed set of potential formulae) could be a topic for future research.

We want to remark that the methodology illustrated in this paper can also be applied in slightly modified settings. It is possible to construct an initial state that explains several observations in the sense that different revision sequences start in the same state. This is reasonable, e.g., when thinking about an expert reasoning about different cases (the initial state representing the expert's background knowledge) or identical copies of software agents being exposed to different situations. Our work is focused on reasoning using observations of other agents, but observing oneself can be useful as well. By keeping an observation of itself an agent may reason about what other agents can conclude about it, which is important when trying to keep certain information secret. The results can also be applied for slight variations of the assumed belief revision framework. For example, it is possible to allow the core belief to be revised or to relax the restriction that new inputs are always appended to the end of ρ in an epistemic state $[\rho, \blacktriangle]$. The interested reader is referred to [24].

Our work has contact points to many other fields in AI research. Most obvious is its relation to belief revision. The intuitive interpretation we

used for the assumed revision framework is incorporation of evidence [13]. However, the representation of the epistemic state as a sequence of formulae does not distinguish between background knowledge and evidence. When applying the results, a more detailed analysis of the intended meaning of the concepts involved and a corresponding interpretation of the results would be needed. Reasoning about other agents is central for many areas, e.g., multi-agent systems, user modelling, goal and plan recognition, etc. Here we investigated one specific aspect. In reasoning about action and change, the question is often to find an action sequence that would cause a particular evolution of the world—either to achieve some goal (planning), or to find out what happened (abduction). Often, the initial state and the effects of an action are specified. In our setting, the effect of a revision input is not quite clear. It might be accepted by the agent or not and beliefs triggered by the input heavily depend on the initial state. Trying to come up with hypotheses about the inner mechanisms of an observed system, which could be interpreted as its initial state that determines its future behaviour, is a topic treated also in induction.

We are not aware of work that investigates reasoning about the evolution of an observed agent's beliefs matching our setting. So we want to conclude by mentioning some papers that investigate similar questions. [11] considers a much richer belief revision framework in a dialogue context. However, the focus is on progressing beliefs through a sequence of speech acts starting in a given initial state of the agents. This and many other publications utilise modal logics for representing agents' beliefs, [14] being another example also handling the *dynamics* of these beliefs. Often there are proof systems or model checkers for the logics presented, but model generation, which is what we are doing in this paper, generally seems to be a problem. This means that if the initial state is not given, hypotheses can only be tested (via proofs) but not systematically generated. However, this is what calculating a potential initial state and the corresponding belief trace is.

The papers [2, 26], which are dealing with update rather than belief revision, start from a sequence of partial descriptions of an evolving world and try to identify preferred trajectories explaining this sequence. [2] intends to sharpen the information about the last state of the world and concentrates on a particular preference relation, giving a representation result. [26] compares different possible preference relations among trajectories, positioning the approach with respect to revision and update. However, both allow for arbitrary changes at any point in time, i.e., they do not allow to integrate information about which actions were performed nor reason about possible outcomes of an action. Recall that although our observation contains the revision input received, this does not mean that it is actually accepted.

Acknowledgments

We thank the editors and anonymous reviewers for very constructive comments that helped to improve an earlier version of this paper. We also want to thank (in lexicographic order) Gerhard Brewka, James P. Delgrande, Didier Dubois, Andreas Herzig, Gabriele Kern-Isberner, Sébastien Konieczny, Jérôme Lang, Wiebe van der Hoek, and Hans van Ditmarsch for helpful discussions on the topic.

References

[1] C. Alchourrón, P. Gärdenfors & D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

[2] S. Berger, D. Lehmann & K. Schlechta. Preferred history semantics for iterated updates. *Journal of Logic and Computation*, 9(6):817–833, 1999.

[3] R. Booth. On the logic of iterated non-prioritised revision. In G. Kern-Isberner, W. Rödder & F. Kulmann, eds., *Conditionals, Information, and Inference. International Workshop, WCII 2002, Hagen, Germany, May 13–15, 2002, Revised Selected Papers*, vol. 3301 of *Lecture Notes in Artificial Intillegence*, pp. 86–107. Springer, 2005.

[4] R. Booth, T. Meyer & K.-S. Wong. A bad day surfing is better than a good day working: How to revise a total preorder. In Doherty et al. [15], pp. 230–238.

[5] R. Booth & A. Nittka. Reconstructing an agent's epistemic state from observations about its beliefs and non-beliefs. *Journal of Logic and Computation*. Forthcoming.

[6] R. Booth & A. Nittka. Beyond the rational explanation. In J. Delgrande, J. Lang, H. Rott & J.-M. Tallon, eds., *Belief Change in Rational Agents: Perspectives from Artificial Intelligence, Philosophy, and Economics*, no. 05321 in Dagstuhl Seminar Proceedings. 2005.

[7] R. Booth & A. Nittka. Reconstructing an agent's epistemic state from observations. In L.P. Kaelbling & A. Saffiotti, eds., *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30–August 5, 2005*, pp. 394–399. Professional Book Center, 2005.

[8] R. Booth & J.B. Paris. A note on the rational closure of knowledge bases with both positive and negative knowledge. *Journal of Logic, Language and Information*, 7(2):165–190, 1998.

[9] C. Boutilier. Revision sequences and nested conditionals. In R. Bajcsy, ed., *Proceedings of the 13th International Joint Conference on Artificial Intelligence. (IJCAI-93) Chambéry, France, August 28–September 3, 1993.*, pp. 519–525. Morgan Kaufmann, 1993.

[10] R.I. Brafman & M. Tennenholtz. Modeling agents as qualitative decision makers. *Artificial Intelligence*, 94(1–2):217–268, 1997.

[11] L.F. del Cerro, A. Herzig, D. Longin & O. Rifi. Belief reconstruction in cooperative dialogues. In F. Giunchiglia, ed., *Artificial Intelligence: Methodology, Systems, and Applications, 8th International Conference, AIMSA '98, Sozopol, Bulgaria, September 21–13, 1998, Proceedings*, vol. 1480 of *Lecture Notes in Computer Science*, pp. 254–266. Springer, 1998.

[12] A. Darwiche & J. Pearl. On the logic of iterated belief revision. *Artificial Intelligence*, 89(1–2):1–29, 1997.

[13] J.P. Delgrande, D. Dubois & J. Lang. Iterated revision as prioritized merging. In Doherty et al. [15], pp. 210–220.

[14] H. van Ditmarsch, W. van der Hoek & B.P. Kooi. *Dynamic Epistemic Logic.*, vol. 337 of *Synthese Library*. Springer-Verlag, 2007.

[15] P. Doherty, J. Mylopoulos & C.A. Welty, eds. *Proceedings, Tenth International Conference on Principles of Knowledge Representation and Reasoning, Lake District of the United Kingdom, June 2–5, 2006.* AAAI Press, 2006.

[16] S.O. Hansson, E. Fermé, J. Cantwell & M. Falappa. Credibility-limited revision. *Journal of Symbolic Logic*, 66(4):1581–1596, 2001.

[17] D. Lehmann. Belief revision, revised. In C.S. Mellish, ed., *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, IJCAI 95, Montréal, Québec, Canada, August 20–25 1995*, pp. 1534–1540. Morgan Kaufmann, 1995.

[18] D. Lehmann & M. Magidor. What does a conditional knowledge base entail? *Artificial Intelligence*, 55(1):1–60, 1992.

[19] D. Makinson. Screened revision. *Theoria*, 63(1–2):14–23, 1997.

[20] A. Nayak. Iterated belief change based on epistemic entrenchment. *Erkenntnis*, 41(3):353–390, 1994.

[21] A. Nayak, M. Pagnucco & P. Peppas. Dynamic belief revision operators. *Artificial Intelligence*, 146(2):193–228, 2003.

[22] B. Nebel. Base revision operations and schemes: Semantics, representation and complexity. In A.G. Cohn, ed., *Proceedings of the Eleventh European Conference on Artificial Intelligence (ECAI-94), Amsterdam, The Netherlands, August 8–12, 1994*, pp. 342–345. John Wiley and Sons, 1994.

[23] A. Nittka. Reasoning about an agent based on its revision history with missing inputs. In M. Fisher, W. van der Hoek, B. Konev & A. Lisitsa, eds., *Logics in Artificial Intelligence, 10th European Conference, JELIA 2006, Liverpool, UK, September 13–15, 2006, Proceedings*, vol. 4160 of *Lecture Notes in Computer Science*, pp. 373–385. Springer, 2006.

[24] A. Nittka. *A Method for Reasoning About Other Agents' Beliefs from Observations*. Ph.D. thesis, Leipzig University, 2008.

[25] O. Papini. Iterated revision operations stemming from the history of an agent's observations. In M.-A. Williams & H. Rott, eds., *Frontiers of Belief Revision*, pp. 279–301. Kluwer Academic Press, 2001.

[26] F.D. de Saint-Cyr & J. Lang. Belief extrapolation (or how to reason about observations and unpredicted change). In D. Fensel, F. Giunchiglia, D.L. McGuinness & M.-A. Williams, eds., *Proceedings of the Eights International Conference on Principles and Knowledge Representation and Reasoning (KR-02), Toulouse, France, April 22–25, 2002*, pp. 497–508. Morgan Kaufmann, 2002.

A Logical Structure for Strategies

R. Ramanujam
Sunil Simon

The Institute of Mathematical Sciences
Central Institutes of Technology (C.I.T.) Campus, Taramani
Chennai 600 113, India

{jam,sunils}@imsc.res.in

Abstract

We consider a logic for reasoning about *composite* strategies in games, where players' strategies are like programs, composed structurally. These depend not only on conditions that hold at game positions but also on properties of other players' strategies. We present an axiomatization for the logic and prove its completeness.

1 Summary

Extensive form turn-based games are trees whose nodes are game positions and branches represent moves of players. With each node is associated a player whose turn it is to move at that game position. A player's *strategy* is then simply a subtree which contains a unique successor for every node where it is this player's turn to make a move, and contains all successors (from the game tree) for nodes where other players make moves. Thus a strategy is an advice function that tells a player what move to play when the game reaches any specific position. In two-player win/loss games, analysis of the game amounts to seeing if either player has a winning strategy from any starting position, and if possible, synthesize such a winning strategy.

In multi-player games where the outcomes are not merely winning and losing, the situation is less clear. Every player has a preference for certain outcomes and hence cooperation as well as conflict become strategically relevant. Moreover, each player has some expectations (and assumptions) about strategies adopted by other players, and fashions her response appropriately. In such situations, game theory tries to explain what *rational* players would do.

In so-called *small* (normal form) games, where the game consists of a small fixed number of moves (often one move chosen independently by each player), strategies have little structure, and prediction of stable behaviour (equilibrium strategy profiles) is possible. However, this not only becomes difficult in games with richer structure and long sequences of moves, it is

Giacomo Bonanno, Wiebe van der Hoek, Michael Wooldridge (eds.). *Logic and the Foundations of Game and Decision Theory (LOFT 7).* Texts in Logic and Games 3, Amsterdam

also less clear how to postulate behaviour of rational players. Moreover, if we look to game theory not only for existence of equilibria but also *advice* to players on how to play, the structure of strategies followed by players becomes relevant.

Even in games of perfect information, if the game structure is sufficiently rich, we need to re-examine the notion of strategy as a function that determines a player's move in every game position. Typically, the game position is itself only *partially known*, in terms of properties that the player can test for. Viewed in this light, strategies are like *programs*, built up systematically from atomic decisions like *if* b *then* a where b is a condition checked by the player to hold (at some game position) and a is a move available to the player at that position.

There is another dimension to strategies, namely that of responses to other players' moves. The notion of each player independently deciding on a strategy needs to be re-examined as well. A player's chosen strategy depends on the player's perception of apparent strategies followed by other players. Even when opponents' moves are visible, an opponent's strategy is not known completely as a function. Therefore the player's strategy is necessarily partial as well.

The central idea of this paper is to suggest that it helps to study *strategies given by their properties*. Hence, assumptions about strategies can be partial, and these assumptions can in turn be structurally built into the specification of other strategies. This leads us to proposing a logical structure for strategies, where we can reason with assertions of the form "(partial) strategy σ *ensures* the (intermediate) condition α".

This allows us to look for *induction principles* which can be articulated in the logic. For instance, we can look at what conditions must be maintained locally (by one move) to influence an outcome eventually. Moreover, we can compare strategies in terms of what conditions they can enforce.

The main contributions of this paper are:

- We consider non-zero-sum games over finite graphs, and consider best response strategies (rather than winning strategies).
- The reasoning carried out works explicitly with the structure of strategies rather than existence of strategies.
- We present a logic with structured strategy specifications and formulas describe how strategies ensure outcomes.
- We present an axiom system for the logic and prove that it is complete.

1.1 Other work

Games are quite popularly used to solve certain decision problems in logic. Probably the best example of a logical game is the Ehrenfeucht-Fraïssé game which is played on two structures to check whether a formula of a certain logic can distinguish between these structures [7]. Games are also used as tools to solve the satisfiability and model checking questions for various modal and temporal logics [12]. Here, an existential and a universal player play on a formula to decide if the formula is satisfiable. The satisfiability problem is then characterised by the question of whether the existential player has a winning strategy in the game. These kinds of games designed specifically for semantic evaluation are generally called *logic games*.

Recently, the advent of computational tasks on the world-wide web and related security requirements have thrown up many game theoretic situations. For example, signing contracts on the web requires interaction between principals who do not know each other and typically distrust each other. Protocols of this kind which involve *selfish agents* can be easily viewed as strategic games of imperfect information. These are complex interactive processes which critically involve players reasoning about each others' strategies to decide on how to act. In this approach, instead of designing games to solve specific logical tasks, one can use logical systems to study structure of games and to reason about them.

Game logics are situated in this context, employing modal logics (in the style of logics of programs) to study logical structure present in games. Parikh's work on propositional game logic [13] initiated the study of game structure using algebraic properties. Pauly [14] has built on this to provide interesting relationships between programs and games, and to describe coalitions to achieve desired goals. Bonnano [5] suggested obtaining game theoretic solution concepts as characteristic formulas in modal logic. Van Benthem [2] uses dynamic logic to describe games as well as strategies. Van Ditmarsch [6] uses a dynamic epistemic language to study complex information change caused by actions in games. The relationship between games defined by game logics and that of logic games, is studied by van Benthem in [3].

On the other hand, the work on Alternating Temporal Logic [1] considers selective quantification over paths that are possible outcomes of games in which players and an environment alternate moves. Here, we talk of the existence of a strategy for a coalition of players to force an outcome. [8] draws parallels between these two lines of work, that of Pauly's coalition logics and alternating temporal logic. It is to be noted that in these logics, the reasoning is about *existence* of strategies, and the strategies themselves do not figure in formulas.

In the work of [10] and [11], van der Hoek and co-authors develop logics for strategic reasoning and equilibrium concepts and this line of work is closest to ours in spirit. Our point of departure is in bringing logical structure into strategies rather than treating strategies as atomic. In particular, the strategy specifications we use are partial (in the sense that a player may assume that an opponent plays a whenever p holds, without knowing under what conditions the opponent's strategy picks another move b), allowing for more generality in reasoning. In the context of programs, logics like propositional dynamic logic [9] explicitly analyse the structure of programs. This approach has been very useful in program verification.

2 Game Arenas

We begin with a description of game models on which formulas of the logic will be interpreted. We use the graphical model for extensive form turn-based multiplayer games, where at most one player gets to move at each game position.

Game arena

Let $N = \{1, 2, \ldots, n\}$ be a non-empty finite set of players and $\Sigma = \{a_1, a_2, \ldots, a_m\}$ be a finite set of action symbols, which represent *moves* of players. A **game arena** is a *finite* graph $\mathcal{G} = (W, \longrightarrow, w_0, \chi)$ where W is the set of nodes which represents the *game positions*, $\longrightarrow : (W \times \Sigma) \to W$ is a function also called the move function, w_0 is the initial node of the game.

Let the set of successors of $w \in W$ be defined as $\vec{w} = \{w' \in W \mid w \xrightarrow{a} w'$ for some $a \in \Sigma\}$. A node w is said to be *terminal* if $\vec{w} = \varnothing$. $\chi : W \to N$ assigns to each node w in W the player who "owns" w: that is, if $\chi(w) = k$ and w is not terminal then player k has to pick a move at w.

In an arena defined as above, the play of a game can be viewed as placing a token on w_0. If player k owns the game position w_0 i.e $\chi(w_0) = k$ and she picks an action 'a' which is enabled for her at w_0, then the new game position moves the token to w' where $w_0 \xrightarrow{a} w'$. A play in the arena is simply a sequence of such moves. Formally, a play in $\mathcal{G}$ is a finite path $\rho = w_0 \xrightarrow{a_1} w_1 \xrightarrow{a_2} \ldots \xrightarrow{a_k} w_k$ where w_k is terminal, or it is an infinite path $\rho = w_0 \xrightarrow{a_1} w_1 \xrightarrow{a_2} \ldots$ where $\forall i : w_i \xrightarrow{a_i} w_{i+1}$ holds. Let *Plays* denote the set of all plays in the arena.

With a game arena $\mathcal{G} = (W, \longrightarrow, w_0, \chi)$, we can associate its *tree unfolding* also referred to as the *extensive form* game tree $\mathcal{T} = (S, \Rightarrow, s_0, \lambda)$ where $(S, \Rightarrow)$ is a countably infinite tree rooted at s_0 with edges labelled by Σ and $\lambda : S \to W$ such that:

- $\lambda(s_0) = w_0$.
- For all $s, s' \in S$, if $s \stackrel{a}{\Longrightarrow} s'$ then $\lambda(s) \xrightarrow{a} \lambda(s')$.

- If $\lambda(s) = w$ and $w \xrightarrow{a} w'$ then there exists $s' \in S$ such that $s \stackrel{a}{\Longrightarrow} s'$ and $\lambda(s') = w'$.

Given the tree unfolding of a game arena $\mathcal{T}$, a node s in it, we can define the *restriction* of $\mathcal{T}$ to s, denoted $\mathcal{T}_s$ to be the subtree obtained by retaining only the unique path from root s_0 to s and the subtree rooted at s.

Games and winning conditions

Let $\mathcal{G}$ be an arena as defined above. The arena merely defines the rules about how the game progresses and terminates. More interesting are *winning conditions*, which specify the game *outcomes*. We assume that each player has a preference relation over the set of plays. Let $\preceq^i \subseteq (\mathsf{Plays} \times \mathsf{Plays})$ be a complete, reflexive, transitive binary relation denoting the preference relation of player i. Then the game G is given as, $G = (\mathcal{G}, \{\preceq^i\}_{i \in N})$.

Then a game is defined as the pair $G = (\mathcal{G}, (\preceq^i)_{i \in N})$.

Strategies

For simplicity we will restrict ourselves to two player games, i.e., $N = \{1, 2\}$. It is easy to extend the notions introduced here to the general case where we have n players.

Let the game graph be represented by $\mathcal{G} = (W^1, W^2, \longrightarrow, s_0)$ where W^1 is the set of positions of player 1, W^2 that of player 2. Let $W = W^1 \cup W^2$.

Let $\mathcal{T}$ be the tree unfolding of the arena and s_1 a node in it. A *strategy* for player 1 at node s_1 is given by: $\mu = (S^1_\mu, S^2_\mu, \Rightarrow_\mu, s_1)$ is a subtree of $\mathcal{T}_{s_1}$ which contains the unique path from root s_0 to s_1 in $\mathcal{T}$ and is the least subtree satisfying the following properties:

- $s_1 \in S^1_\mu$, where $\chi(\lambda(s_1)) = 1$.
- For every s in the subtree of $\mathcal{T}_G$ rooted at s_1,
  - if $s \in S^1_\mu$ then for some $a \in \Sigma$, for each s' such that $s \stackrel{a}{\Longrightarrow} s'$, we have $s \stackrel{a}{\Longrightarrow}_\mu s'$.
  - if $s \in S^2_\mu$, then for every $b \in \Sigma$, for each s' such that $s \stackrel{b}{\Longrightarrow} s'$, we have $s \stackrel{b}{\Longrightarrow}_\mu s'$.

Let Ω_i denote the set of all strategies of Player i in G, for $i = 1, 2$. A strategy profile $\langle \mu, \tau \rangle$ defines a unique play ρ^τ_μ in the game $\mathcal{G}$.

3 The Logic

We now present a logic for reasoning about composite strategies. The syntax of the logic is presented in two layers, that of *strategy specification* and *game formulas*.

Atomic strategy formulas specify, for a player, what conditions she tests for before making a move. Since these are intended to be bounded memory strategies, the conditions are stated as *past time* formulas of a simple tense logic. Composite strategy specifications are built from atomic ones using connectives (without negation). We crucially use an implication of the form: "if the opponent's play conforms to a strategy π then play σ".

Game formulas describe the game arena in a standard modal logic, and in addition specify the result of a player following a particular strategy at a game position, to choose a specific move a, to *ensure* an intermediate outcome α . Using these formulas one can specify how a strategy helps to eventually *win* an outcome α.

Before we describe the logic and give its semantics, some prelimiaries will be useful. Below, for any countable set X, let $\mathsf{Past}(X)$ be a set of formulas given by the following syntax:

$$\psi \in \mathsf{Past}(X) := x \in X \mid \neg\psi \mid \psi_1 \vee \psi_2 \mid \diamondsuit\psi.$$

Such past formulas can be given meaning over finite sequences. Given any sequence $\xi = t_0 t_1 \cdots t_m$, $V : \{t_0, \cdots, t_m\} \to 2^X$, and k such that $0 \leq k \leq m$, the truth of a past formula $\psi \in \mathsf{Past}(X)$ at k, denoted $\xi, k \models \psi$ can be defined as follows:

- $\xi, k \models p$ iff $p \in V(t_k)$.
- $\xi, k \models \neg\psi$ iff $\xi, k \not\models \psi$.
- $\xi, k \models \psi_1 \vee \psi_2$ iff $\xi, k \models \psi_1$ or $\xi, k \models \psi_2$.
- $\xi, k \models \diamondsuit\psi$ iff there exists a $j : 0 \leq j \leq k$ such that $\xi, j \models \psi$.

Strategy specifications

For simplicity of presentation, we stick with two player games, where the players are Player 1 and Player 2. Let $\overline{i} = 2$ when $i = 1$ and $\overline{i} = 1$ when $i = 2$.

Let $P^i = \{p^i_0, p^i_1, \ldots\}$ be a countable set of proposition symbols where $\tau_i \in P_i$, for $i \in \{1, 2\}$. Let $P = P^1 \cup P^2 \cup \{\mathsf{leaf}\}$. τ_1 and τ_2 are intended to specify, at a game position, which player's turn it is to move. The proposition leaf specifies whether the position is a terminal node.

Further, the logic is parametrized by the finite alphabet set $\Sigma = \{a_1, a_2, \ldots, a_m\}$ of players' moves and we only consider game arenas over Σ.

Let $\mathsf{Strat}^i(P^i)$, for $i = 1, 2$ be the set of strategy specifications given by the following syntax:

$$\mathsf{Strat}^i(P^i) := [\psi \mapsto a_k]^i \mid \sigma_1 + \sigma_2 \mid \sigma_1 \cdot \sigma_2 \mid \pi \Rightarrow \sigma$$

where $\pi \in \mathsf{Strat}^{\overline{i}}(P^1 \cap P^2)$, $\psi \in \mathsf{Past}(P^i)$ and $a_k \in \Sigma$.

The idea is to use the above constructs to specify properties of strategies. For instance the interpretation of a player i specification $[p \mapsto a]^i$ will be to choose move "a" for every i node where p holds. $\pi \Rightarrow \sigma$ would say, at any node player i sticks to the specification given by σ if on the history of the play, all moves made by $\overline{i}$ conforms to π. In strategies, this captures the aspect of players actions being responses to the opponents moves. As the opponents complete strategy is not available, the player makes a choice taking into account the apparent behaviour of the opponent on the history of play.

For a game tree $\mathcal{T}$, a node s and a strategy specification $\sigma \in \mathsf{Strat}^i(P^i)$, we define $\mathcal{T}_s \restriction \sigma = (S_\sigma, \Longrightarrow_\sigma, s_0)$ to be the least subtree of $\mathcal{T}_s$ which contains $\rho^s_{s_0}$ (the unique path from s_0 to s) and closed under the following condition.

- For every s' in S_σ such that $s \Longrightarrow^*_\sigma s'$,
  - s' is an i node: $s' \overset{a}{\Longrightarrow} s''$ and $a \in \sigma(s') \Leftrightarrow s' \overset{a}{\Longrightarrow}_\sigma s''$.
  - s' is an $\overline{i}$ node: $s' \overset{a}{\Longrightarrow} s'' \Leftrightarrow s' \overset{a}{\Longrightarrow}_\sigma s''$.

Given a game tree $\mathcal{T}$ and a node s in it, let $\rho^s_{s_0} : s_0 \overset{a_1}{\Longrightarrow} s_1 \cdots \overset{a_m}{\Longrightarrow} s_m = s$ denote the unique path from s_0 to s. For a strategy specification $\sigma \in \mathsf{Strat}^i(P^i)$ and a node s we define $\sigma(s)$ as follows:

- $[\psi \mapsto a]^i(s) = \begin{cases} \{a\} & \text{if } s \in W^i \text{ and } \rho^s_{s_0}, m \models \psi \\ \Sigma & \text{otherwise} \end{cases}$
- $(\sigma_1 + \sigma_2)(s) = \sigma_1(s) \cup \sigma_2(s)$.
- $(\sigma_1 \cdot \sigma_2)(s) = \sigma_1(s) \cap \sigma_2(s)$.
- $(\pi \Rightarrow \sigma)(s) = \begin{cases} \sigma(s) & \text{if } \forall j : 0 \leq j < m,\ a_j \in \pi(s_j) \\ \Sigma & \text{otherwise} \end{cases}$

We say that a path $\rho^{s'}_s : s = s_1 \overset{a_1}{\Longrightarrow} s_2 \cdots \overset{a_{m-1}}{\Longrightarrow} s_m = s'$ in $\mathcal{T}$ *conforms* to σ if $\forall j : 1 \leq j < m,\ a_j \in \sigma(s_j)$. When the path constitutes a proper play, i.e., when $s = s_0$, we say that the play conforms to σ.

Syntax

The syntax of the logic is given by:

$$\Pi := p \in P \mid \neg\alpha \mid \alpha_1 \vee \alpha_2 \mid \langle a \rangle \alpha \mid \langle \overline{a} \rangle \alpha \mid \diamondsuit \alpha \mid (\sigma)_i : c \mid \sigma \rightsquigarrow_i \beta$$

where $c \in \Sigma$, $\sigma \in \mathsf{Strat}^i(P^i)$, $\beta \in \mathsf{Past}(P^i)$. The derived connectives $\wedge$, $\supset$ and $[a]\alpha$ are defined as usual. Let $\diamondsuit \alpha = \neg \boxdot \neg \alpha$, $\langle N \rangle \alpha = \bigvee_{a \in \Sigma} \langle a \rangle \alpha$, $[N]\alpha = \neg \langle N \rangle \neg \alpha$, $\langle P \rangle \alpha = \bigvee_{a \in \Sigma} \langle \overline{a} \rangle \alpha$ and $[P] = \neg \langle P \rangle \neg \alpha$.

The formula $(\sigma)_i : c$ asserts, at any game position, that the strategy specification σ for player i suggests that the move c can be played at that position. The formula $\sigma \rightsquigarrow_i \beta$ says that from this position, there is a way of following the strategy σ for player i so as to ensure the outcome β. These two modalities constitute the main constructs of our logic.

Semantics

The models for the logic are extensive form game trees along with a valuation function. A model $M = (\mathcal{T}, V)$ where $\mathcal{T} = (S^1, S^2, \longrightarrow, s_0)$ is a game tree as defined in Section 2, and $V : S \to 2^P$ is the valuation function, such that:

- For $i \in \{1, 2\}$, $\tau_i \in V(s)$ iff $s \in S^i$.
- $\mathsf{leaf} \in V(s)$ iff $\mathsf{moves}(s) = \varphi$.

where for any node s, $\mathsf{moves}(s) = \{a \mid s \stackrel{a}{\Longrightarrow} s'\}$.

The truth of a formula $\alpha \in \Pi$ in a model M and position s (denoted $M, s \models \alpha$) is defined by induction on the structure of α, as usual. Let $\rho^s_{s_0}$ be $s_0 \stackrel{a_0}{\Longrightarrow} s_1 \cdots \stackrel{a_{m-1}}{\Longrightarrow} s_m = s$.

- $M, s \models p$ iff $p \in V(s)$.
- $M, s \models \neg\alpha$ iff $M, s \not\models \alpha$.
- $M, s \models \alpha_1 \vee \alpha_2$ iff $M, s \models \alpha_1$ or $M, s \models \alpha_2$.
- $M, s \models \langle a \rangle \alpha$ iff there exists $s' \in W$ such that $s \stackrel{a}{\rightarrow} s'$ and $M, s' \models \alpha$.
- $M, s \models \langle \overline{a} \rangle \alpha$ iff $m > 0$, $a = a_{m-1}$ and $M, s_{m-1} \models \alpha$.
- $M, s \models \diamondsuit \alpha$ iff there exists $j : 0 \leq j \leq m$ such that $M, s_j \models \alpha$.
- $M, s \models (\sigma)_i : c$ iff $c \in \sigma(s)$.
- $M, s \models \sigma \rightsquigarrow_i \beta$ iff for all s' in $\mathcal{T}_s \restriction \sigma$ such that $s \Longrightarrow^* s'$, we have $M, s' \models \beta \wedge (\tau_i \supset \mathsf{enabled}_\sigma)$,

where $\mathsf{enabled}_\sigma \equiv \bigvee_{a \in \Sigma} (\langle a \rangle \mathsf{True} \wedge (\sigma)_i : a)$.

The notions of satisfiablility and validity can be defined in the standard way. A formula α is *satisfiable* iff there exists a model M, there exists s such that $M, s \models \alpha$. A formula α is said to be *valid* iff for all models M, for all s, we have $M, s \models \alpha$.

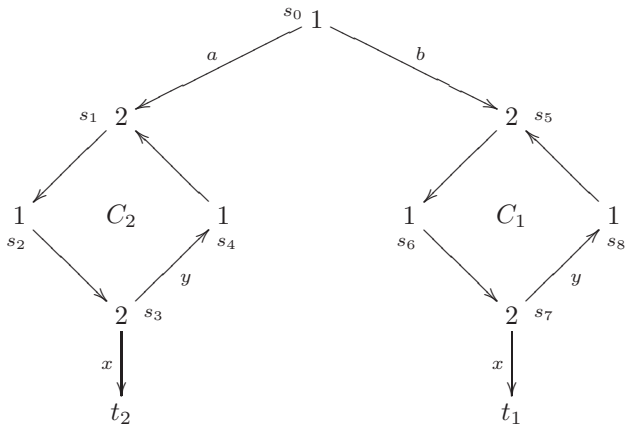


FIGURE 1.

4 Example

Probably the best way to illustrate the notion of strategy specification is to look at heuristics used in large games like chess, go, checkers, etc. A heuristic strategy is basically a partial specification, since it involves checking local properties like patterns on the board and specifying actions when certain conditions hold. For instance, a typical strategy specification for chess would be of the form:

- If a pawn double attack is possible then play the action resulting in the fork.

Note that the above specification is in contrast with a specific advice of the form:

- If a pawn is on f2 and the opponent rook and knight are on e5 and g5 respectively then move f2–f4.

A strategy would prescribe such specific advice rather than a generic one based on abstract game position properties. Heuristics are usually employed when the game graph being analysed is too huge for a functional strategy to be specified. However, we refrain from analysing chess here due to the difficulty in formally presenting the game arena and the fact that it fails to give much insight into the working of our logic. Below we look at a few simple examples which illustrates the logic.

Example 4.1. Consider the game shown in Figure 1. Players alternate moves with 1 starting at s_0. There are two cycles $C_1 : s_5 \to s_6 \to s_7 \to$

$s_8 \to s_5$, $C_2 : s_1 \to s_2 \to s_3 \to s_4 \to s_1$ and two terminal nodes t_1 and t_2. Let the preference ordering of player 1 be $t_1 \preceq^1 t_2 \preceq^1 C_2 \preceq^1 C_1$. As far as player 2 is concerned $t_1 \preceq^2 C_1$ and he is indifferent between C_2 and t_2. However, he prefers C_2 or t_2 over $\{C_1, t_1\}$. Equilibrium reasoning will advise player 1 to choose the action "b" at s_0 since at position s_7 it is irrational for 2 to move x as it will result in 2's worst outcome. However the utility difference between C_1 and t_1 for 2 might be negligible compared to the incentive of staying in the "left" path. Therefore 2 might decide to punish 1 for moving b when 1 knew that $\{C_2, t_2\}$ was equally preferred by 2. Even though t_1 is the worst outcome, at s_7 player 2 can play x to implement the punishment. Let $V(p_j) = \{s_3, s_7\}, V(p_{\mathsf{init}}) = \{s_0\}, V(p_{\mathsf{good}}) = \{s_0, s_1, s_2, s_3, s_4\}$ and $V(p_{\mathsf{punish}}) = \{s_0, s_5, s_6, s_7, t_1\}$. The local objective of 2 will be to remain on the good path or to implement the punishment. Player 2 strategy specification can be written as

$$\pi \equiv ([p_{\mathsf{init}} \mapsto b]^1 \Rightarrow [p_j \mapsto x]^2) \cdot ([p_{\mathsf{init}} \mapsto a]^1 \Rightarrow [p_j \mapsto y]^2).$$

We get that $\pi \rightsquigarrow_2 (p_{\mathsf{good}} \vee p_{\mathsf{punish}})$. Player 1, if he knows 2's strategy, might be tempted to play "a" at s_0 by which the play will end up in C_2. Let the proposition p_{worst} hold at t_1 which is the worst outcome for player 1. Then we have $[p_{\mathsf{init}} \mapsto a]^1 \rightsquigarrow_1 \neg p_{\mathsf{worst}}$. This says that if player 1 chooses a at the initial position then he can ensure that the worst outcome is avoided.

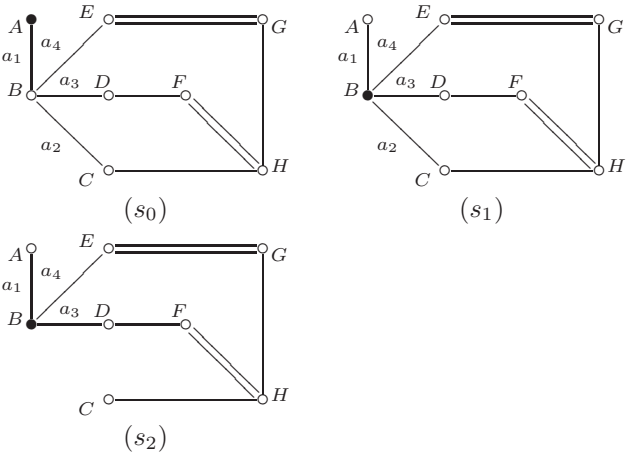


FIGURE 2. Sabotage Game.

Example 4.2. The sabotage game [4] is a two player zero sum game where one player moves along the edges of a labelled graph and the other player

removes an edge in each round. Formally let a Σ labelled graph R for some alphabet set Σ is $R = (V, e)$ where V is the set of vertices and $e : V \times \Sigma \to V$ is the edge function. The sabotage game is played as follows: initially we consider the graph $R_0 = (V_0, e_0, v_0)$. There are two players, *Runner* and *Blocker* who move alternatingly where the *Runner* starts the run from vertex v_0. In round n, the *Runner* moves one step further along an existing edge of the graph. I.e., he chooses a vertex $v_{n+1} \in V$ such that there exists some $a \in \Sigma$ with $e_n(v_n, a) = v_{n+1}$. Afterwards the *Blocker* removes one edge of the graph. i.e., he chooses two vertices u and v such that for some $a \in \Sigma$, $e_n(u, a) = v$ and defines the edge function e_{n+1} to be same as that of e_n except that $e_{n+1}(u, a)$ will not be defined. The graph $R_{n+1} = (V, e_{n+1}, v_{n+1})$. We can have a reachability condition as the winning condition. I.e., the *Runner* wins iff he can reach a given vertex called the goal. The game ends, if either the *Runner* gets stuck or if the winning condition is satisfied.

It is easy to build a conventional game arena for the sabotage game where player positions alternate. The game arena will have as its local states subgraphs of R with the current position of *Runner* indicated. I.e., $W = \mathsf{Edges} \times V$ where Edges is the set of all partial edge functions $e : V \times \Sigma \to V$. Let W^1 and W^2 be the set of game positions for *Runner* and *Blocker* respectively. The initial vertex $s_0 = (e_0, v_0)$ and $s_0 \in W^1$. Let $s = (e, v)$ and $s' = (e', v')$ be any two nodes in the arena. The transition is defined as follows.

- if $s \in W^1$ and $e(v, a) = v'$ then $s \xrightarrow{a} s'$, $e = e'$ and $s' \in W^2$
- if $s \in W^2$, for some $u, u' \in W$ we have $e(u, a) = u'$ and e' is the same as e except that $e'(u, a)$ is not defined, then $s \xrightarrow{(u,a,u')} s'$, $v = v'$ and $s' \in W^1$.

Figure 2 shows the first three game positions in a possible run of the sabotage game. The game starts with *Runner* moving from node A to node B. The blocker then removes the edge a_2 adjacent to node B and the game continues. In the formulas given below, we will refer to *Runner* as player 1 and *Blocker* as player 2.

Since the Runner's objective is to not get stuck, he might reason as follows. If it is the case that the $\mathsf{Blocker}$ always removes an edge adjacent to the node that Runner has currently selected then try to move to a node which has multiple outgoing edges. We use the following propositions:

- $\mathsf{present}_v$: denotes that the current node of runner is v
- adj_m: denotes that the adjacent node has multiple edges

Let r_v denote the action which removes an adjacent edge of v and $\mathsf{move}_{\mathsf{adj}}$ denote the action which moves to the adjacent node with multiple edges. The Runner's specification can be given as:

- $[\mathsf{present}_v \mapsto r_v]^2 \Rightarrow [\mathsf{adj}_m \mapsto \mathsf{move}_{\mathsf{adj}}]^1$

Consider the situation where all the nodes in the graph have a single outgoing edge and the goal state is a single state. It is quite easy to show that in such a game, the Runner wins iff the start node is the goal or if there is an edge connecting the start node with the goal. This property can be captured by the following proposition:

- g^B_{nice}: denotes that in the graph the start node is not the goal and there is no single edge between start and goal nodes. In other words the graph is "nice" for Blocker.
- adj^R_g: denotes that the Runner's current node is one adjacent to the goal node.

Let r^g_{adj} denote the action which removes the edge connecting the current node of Runner with the goal. Consider the following formula:

- $[(g^B_{\mathsf{nice}} \wedge \mathsf{adj}^R_g) \mapsto r^g_{\mathsf{adj}}]^2 \rightsquigarrow_2 (\mathsf{leaf} \supset \mathsf{win})$

This says that if the graph is "nice" for Blocker and if the current selected node of Runner is one adjacent to the goal then remove the only edge connecting it with the goal. In all the other cases the Blocker can remove any random edge, and so this need not be mentioned in the strategy specification. This specification ensures that when the terminal node is reached then it is winning for Blocker.

5 Axiom System

We now present our axiomatization of the valid formulas of the logic. Before we present the axiomatization, we will find some abbreviations useful:

- $\mathsf{root} = \neg\langle P\rangle\mathsf{True}$ defines the root node to be one that has no predecessors.
- $\delta^\sigma_i(a) = \tau_i \wedge (\sigma)_i : a$ denotes that move "a" is enabled by σ at an i node.
- $\mathsf{inv}^\sigma_i(a, \beta) = (\tau_i \wedge (\sigma)_i : a) \supset [a](\sigma \rightsquigarrow_i \beta)$ denotes the fact that after an "a" move by player i which conforms to σ, $\sigma \rightsquigarrow_i \beta$ continues to hold.
- $\mathsf{inv}^\sigma_{\overline{i}}(\beta) = \tau_{\overline{i}} \supset [N](\sigma \rightsquigarrow_i \beta)$ says that after any move of $\overline{i}$, $\sigma \rightsquigarrow_i \beta$ continues to hold.
- $\mathsf{conf}_\pi = \boxminus(\langle\overline{a}\rangle\tau_{\overline{i}} \supset \langle\overline{a}\rangle(\pi)_{\overline{i}} : a)$ denotes that all opponent moves in the past conform to π.

The axiom schemes

(A0) All the substitutional instances of the tautologies of propositional calculus.

(A1) (a) $[a](\alpha_1 \supset \alpha_2) \supset ([a]\alpha_1 \supset [a]\alpha_2)$

(b) $[\overline{a}](\alpha_1 \supset \alpha_2) \supset ([\overline{a}]\alpha_1 \supset [\overline{a}]\alpha_2)$

(A2) (a) $\langle a\rangle\alpha \supset [a]\alpha$

(b) $\langle \overline{a}\rangle\alpha \supset [\overline{a}]\alpha$

(c) $\langle \overline{a}\rangle \mathsf{True} \supset \neg\langle \overline{b}\rangle \mathsf{True}$ for all $b \neq a$

(A3) (a) $\alpha \supset [a]\langle \overline{a}\rangle\alpha$

(b) $\alpha \supset [\overline{a}]\langle a\rangle\alpha$

(A4) (a) $\diamond\!\!\!-\, \mathsf{root}$

(b) $\boxminus\alpha \equiv (\alpha \wedge [P]\boxminus\alpha)$

(A5) (a) $([\psi \mapsto a]^i)_i : a$ for all $a \in \Sigma$

(b) $\tau_i \wedge ([\psi \mapsto a]^i)_i : c \equiv \neg\psi$ for all $a \neq c$

(A6) (a) $(\sigma_1 + \sigma_2)_i : c \equiv \sigma_1 : c \vee \sigma_2 : c$

(b) $(\sigma_1 \cdot \sigma_2)_i : c \equiv \sigma_1 : c \wedge \sigma_2 : c$

(c) $(\pi \Rightarrow \sigma)_i : c \equiv \mathsf{conf}_\pi \supset (\sigma)_i : c$

(A7) $\sigma \rightsquigarrow_i \beta \supset (\beta \wedge \mathsf{inv}_i^\sigma(a, \beta) \wedge \mathsf{inv}_{\overline{i}}^\sigma(\beta) \wedge (\neg\mathsf{leaf} \supset \mathsf{enabled}_\sigma))$

Inference rules

$$\frac{\alpha, \quad \alpha \supset \beta}{\beta} \quad (MP) \qquad \frac{\alpha}{[a]\alpha} \quad (NG) \qquad \frac{\alpha}{[\overline{a}]\alpha} \quad (\mathsf{NG}\text{-})$$

$$\frac{\alpha \supset [P]\alpha}{\alpha \supset \boxminus\alpha} \qquad (\mathsf{Ind\text{-}past})$$

$$\frac{\alpha \wedge \delta_i^\sigma(a) \supset [a]\alpha, \quad \alpha \wedge \tau_{\overline{i}} \supset [N]\alpha, \quad \alpha \wedge \neg\mathsf{leaf} \supset \mathsf{enabled}_\sigma, \quad \alpha \supset \beta}{\alpha \supset \sigma \rightsquigarrow_i \beta} \qquad (\mathsf{Ind} \rightsquigarrow)$$

The axioms are mostly standard. After the Kripke axioms for the $\langle a\rangle$ modalities, we have axioms that ensure determinacy of both $\langle a\rangle$ and $\langle \overline{a}\rangle$ modalities, and an axiom to assert the uniqueness of the latter. We then have axioms that relate the previous and next modalities with each other, as well as to assert that the past modality steps through the $\langle \overline{a}\rangle$ modality. An

axiom asserts the existence of the root in the past. The rest of the axioms describe the semantics of strategy specifications.

The rule Ind-past is standard, while Ind $\rightsquigarrow$ illustrates the new kind of reasoning in the logic. It says that to infer that the formula $\sigma \rightsquigarrow_i \beta$ holds in all reachable states, β must hold at the asserted state and

- for a player i node after every move which conforms to σ, β continues to hold.
- for a player $\overline{i}$ node after every enabled move, β continues to hold.
- player i does not get stuck by playing σ.

To see the soundness of (A7), suppose it is not valid. Then there exists a node s such that $M, s \models \sigma \rightsquigarrow_i \beta$ and one of the following holds:

- $M, s \not\models \beta$: In this case, from semantics we get that $M, s \not\models \sigma \rightsquigarrow_i \beta$ which is a contradiction.
- $M, s \not\models \mathsf{inv}_i^\sigma(a, \beta)$: In this case, we have $s \in W^i$, $M, s \models (\sigma)_i : a$ and $M, s' \not\models \sigma \rightsquigarrow_i \beta$ where $s \stackrel{a}{\Longrightarrow} s'$. This implies that there is a path $\rho_{s'}^{s_k}$ which conforms to σ and either $M, s_k \not\models \beta$ or $\mathsf{moves}(s_k) \cap \sigma(s_k) = \varphi$. But since $s \stackrel{a}{\Longrightarrow} s'$, we have $\rho_s^{s_k}$ conforms to σ as well. From which it follows that $M, s \not\models \sigma \rightsquigarrow_i \beta$ which is a contradiction.
- $M, s \not\models \mathsf{inv}_{\overline{i}}^\sigma(\beta)$: We have a similar argument as above.
- $M, s \not\models \neg\mathsf{leaf} \supset \mathsf{enabled}_\sigma$: This means that $M, s \models \neg\mathsf{leaf}$ and $M, s \not\models \mathsf{enabled}_\sigma$. Therefore $\mathsf{moves}(s) \cap \sigma(s) = \varnothing$ and by semantics we have $M, s \not\models \sigma \rightsquigarrow_i \beta$ which is a contradiction.

To show that the induction rule preserves validity, suppose that the premise is valid and the conclusion is not. Then for some node s we have $M, s \models \alpha$ and $M, s \not\models \sigma \rightsquigarrow_i \beta$. i.e., there is a path $\rho_s^{s_k}$ which conforms to σ such that $M, s_k \not\models \beta$ or s_k is a non-leaf node and $\sigma(s_k) \cap \mathsf{moves}(s_k) = \varphi$. Let $\rho_s^{s_k}$ be the shortest of such paths.

Suppose $M, s_k \not\models \beta$, then we have the following two cases to consider.

- $s_{k-1} \in W^i$: By assumption on the path $\rho_s^{s_k}$, we have $M, s_{k-1} \models \alpha \wedge \delta_i^\sigma(a_{k-1})$. From validity of $\alpha \supset \beta$(the premise), we have $M, s_k \not\models \alpha$, which implies $M, s_{k-1} \not\models [a_{k-1}]\alpha$. Therefore we get $M, s_{k-1} \not\models (\alpha \wedge \delta_i^\sigma(a_{k-1})) \supset [a_{k-1}]\alpha$, which gives us a contradiction to the validity of a premise.
- $s_{k-1} \in W^{\overline{i}}$: By assumption on the path $\rho_s^{s_k}$, we have $M, s_{k-1} \models \alpha \wedge \tau_{\overline{i}}$. Using an argument similar to the previous case we also get $M, s_{k-1} \not\models [a_{k-1}]\alpha$. Therefore we have $M, s_{k-1} \not\models (\alpha \wedge \tau_{\overline{i}}) \supset [N]\alpha$, giving us a contradiction to the validity of a premise.

If s_k is a non-leaf node and $\sigma(s_k) \cap \mathsf{moves}(s_k) = \varphi$ then we have $M, s_k \models \alpha \wedge \neg\mathsf{leaf}$ and $M, s_k \not\models \mathsf{enabled}_\sigma$. Therefore $M, s_k \not\models (\alpha \wedge \neg\mathsf{leaf}) \supset \mathsf{enabled}_\sigma$, which is the required contradiction.

6 Completeness

To show completeness, we prove that every consistent formula is satisfiable. Let α_0 be a consistent formula, and let W denote the set of all maximal consistent sets (MCS). We use w, w' to range over MCS's. Since α_0 is consistent, there exists an MCS w_0 such that $\alpha_0 \in w_0$.

Define a transition relation on MCS's as follows: $w \xrightarrow{a} w'$ iff $\{\langle a\rangle\alpha | \alpha \in w'\} \subseteq w$. We will find it useful to work not only with MCS's, but also with sets of subformulas of α_0. For a formula α let $\mathsf{CL}(\alpha)$ denote the subformula closure of α. In addition to the usual downward closure, we also require that $\diamondsuit\mathsf{root}, \mathsf{leaf} \in \mathsf{CL}(\alpha)$ and $\sigma \rightsquigarrow_i \beta \in \mathsf{CL}(\alpha)$ implies that $\beta, \mathsf{inv}_i^\sigma(a, \beta), \mathsf{inv}_i^\sigma(\beta), \mathsf{enabled}_\sigma \in \mathsf{CL}(\alpha)$. Let $\mathcal{AT}$ denote the set of all maximal consistent subsets of $\mathsf{CL}(\alpha_0)$, reffered to as *atoms*. Each $t \in \mathcal{AT}$ is a finite set of formulas, we denote the conjunction of all formulas in t by $\widehat{t}$. For a nonempty subset $X \subseteq \mathcal{AT}$, we denote by $\widetilde{X}$ the disjunction of all $\widehat{t}, t \in X$. Define a transition relation on $\mathcal{AT}$ as follows: $t \xrightarrow{a} t'$ iff $\widehat{t} \wedge \langle a\rangle\widehat{t'}$ is consistent. Call an atom t a *root atom* if there does not exist any atom t' such that $t' \xrightarrow{a} t$ for some a. Note that $t_0 = w_0 \cap \mathsf{CL}(\alpha_0) \in \mathcal{AT}$.

Proposition 6.1. There exist $t_1, \ldots, t_k \in \mathcal{AT}$ and $a_1, \ldots a_k \in \Sigma$ $(k \geq 0)$ such that $t_k \xrightarrow{a_k} t_{k-1} \ldots \xrightarrow{a_1} t_0$, where t_k is a root atom.

Proof. Consider the least set R containing t_0 and closed under the following condition: if $t_1 \in R$ and for some $a \in \Sigma$ there exists t_2 such that $t_2 \xrightarrow{a} t_1$, then $t_2 \in R$. Now, if there exists an atom $t' \in R$ such that t' is a root then we are done. Suppose not, then we have $\vdash \widetilde{R} \supset \neg\mathsf{root}$. But then we can show that $\vdash \widetilde{R} \supset [P]\widetilde{R}$. By rule $\mathsf{Ind\text{-}past}$ and above we get $\vdash \widetilde{R} \supset \boxminus\neg\mathsf{root}$. But then $t_0 \in R$ and hence $\vdash \widehat{t_0} \supset \widetilde{R}$ and therefore we get $\vdash \widehat{t_0} \supset \boxminus\neg\mathsf{root}$, contradicting axiom (A4a). Q.E.D.

Above, we have additional properties: for any formula $\diamondsuit\alpha \in t_k$, we also have $\alpha \in t_k$. Further, for all $j \in \{0, \cdots, k\}$, if $\diamondsuit\alpha \in t_j$, then there exists i such that $k \geq i \geq j$ and $\alpha \in t_i$. Both these properties are ensured by axiom (A4b). A detailed proof can be found in the Appendix, Lemma A.2.

Hence it is easy to see that there exist MCS's $w_1, \ldots, w_k \in W$ and $a_1, \ldots a_k \in \Sigma$ $(k \geq 0)$ such that $w_k \xrightarrow{a_k} w_{k-1} \ldots \xrightarrow{a_1} w_0$, where $w_j \cap \mathsf{CL}(\alpha_0) = t_j$. Now this path defines a (finite) tree $T_0 = (S_0, \Longrightarrow_0, s_0)$ rooted at s_0, where $S_0 = \{s_0, s_1, \ldots, s_k\}$, and for all $j \in \{0, \cdots, k\}$, s_j is labelled by the MCS w_{k-j}. The relation $\Longrightarrow_0$ is defined in the obvious manner. From now we will simply say $\alpha \in s$ where s is the tree node, to mean that $\alpha \in w$ where w is the MCS associated with node s.

Inductively assume that we have a tree $T_k = (S_k, \Longrightarrow_k, s_0)$ such that the past formulas at every node have "witnesses" as above. Pick a node $s \in S_k$ such that $\langle a \rangle \mathsf{True} \in s$ but there is no $s' \in S_k$ such that $s \stackrel{a}{\Longrightarrow} s'$. Now, if w is the MCS associated with node s, there exists an MCS w' such that $w \stackrel{a}{\longrightarrow} w'$. Pick a new node $s' \notin S_k$ and define $T_{k+1} = S_k \cup \{s'\}$ and $\Longrightarrow_{k+1} = \Longrightarrow_k \cup \{(s, a, s')\}$, where w' is the MCS associated with s'. It is easy to see that every node in T_{k+1} has witnesses for past formulas as well.

Now consider $T = (S, \Longrightarrow, s_0)$ defined by: $S = \bigcup_{k \geq 0} S_k$ and $\Longrightarrow = \bigcup_{k \geq 0} \Longrightarrow_k$. Define the model $M = (T, V)$ where $V(s) = w \cap P$, where w is the MCS associated with s.

Lemma 6.2. For any $s \in S$, we have the following properties.

1. if $[a]\alpha \in s$ and $s \stackrel{a}{\Longrightarrow} s'$ then $\alpha \in s'$.
2. if $\langle a \rangle \alpha \in s$ then there exists s' such that $s \stackrel{a}{\Longrightarrow} s'$ and $\alpha \in s'$.
3. if $[\overline{a}]\alpha \in s$ and $s' \stackrel{a}{\Longrightarrow} s$ then $\alpha \in s'$.
4. if $\langle \overline{a} \rangle \alpha \in s$ then there exists s' such that $s' \stackrel{a}{\Longrightarrow} s$ and $\alpha \in s'$.
5. if $\boxminus \alpha \in s$ and $s' \Longrightarrow^* s$ then $\alpha \in s'$.
6. if $\diamondminus \alpha \in s$ then there exists s' such that $s' \Longrightarrow^* s$ and $\alpha \in s'$.

Proof. Cases (1) to (5) can be shown using standard modal logic techniques. Case (6) follows from the existence of a root atom (Proposition 6.1) and axiom (A4b). Q.E.D.

Lemma 6.3. For all $\psi \in \mathsf{Past}(P)$, for all $s \in S$, $\psi \in s$ iff $\rho_s, s \models \psi$.

Proof. This follows from Lemma 6.2 using an inductive argument. Q.E.D.

Lemma 6.4. For all i, for all $\sigma \in \mathsf{Strat}^i(P^i)$, for all $c \in \Sigma$, for all $s \in S$, $(\sigma)_i : c \in s$ iff $c \in \sigma(s)$.

Proof. The proof is by induction on the structure of σ. The nontrivial cases are as follows:

$\sigma \equiv [\psi \mapsto a]$:

($\Rightarrow$) Suppose $([\psi \mapsto a]^i)_i : c \in s$. If $c = a$ then the claim holds trivially. If $c \neq a$ then from (A5a) we get that $\neg\psi \in s$, from Lemma 6.3 $\rho_s, s \not\models \psi$. Therefore by definition we have $[\psi \mapsto a]^i(s) = \Sigma$ and $c \in \sigma(w)$.

($\Leftarrow$) Conversely, suppose $([\psi \mapsto a]^i)_i : c \notin s$. From (A5a) we have $a \neq c$. From (A5b) we get $\psi \in s$. By Lemma 6.3 $\rho_s, s \models \psi$. Therefore $c \notin \sigma(s)$ by definition.

$\sigma \equiv \pi \Rightarrow \sigma'$: Let $\rho^s_{s_0} : s_0 \xrightarrow{a_0} \cdots \xrightarrow{a_{k-1}} s_k = s$ be the unique path from the root to s.

($\Rightarrow$) Suppose $(\pi \Rightarrow \sigma')_i : c \in s$. To show $c \in (\pi \Rightarrow \sigma')(s)$. Suffices to show that $\rho^s_{s_0}$ conforms to π implies $c \in \sigma'(s)$. From (A6c) we have $\mathsf{conf}_\pi \supset (\sigma')_i : c \in s$. Rewriting this we get $\Diamond(\langle\overline{a}\rangle\tau_{\overline{i}} \wedge [\overline{a}](\neg(\pi)_{\overline{i}} : a)) \vee (\sigma')_i : c \in s$. We have two cases,

- if $(\sigma')_i : c \in s$ then by induction hypothesis we get $c \in \sigma'(s)$. Therefore by definition $c \in (\pi \Rightarrow \sigma)_i(s)$.
- otherwise we have $\Diamond(\langle\overline{a}\rangle\tau_{\overline{i}} \wedge [\overline{a}](\neg(\pi)_{\overline{i}} : a)) \in s$. From Lemma 6.2(6), there exists $s_l \in \rho_s$ such that $\langle\overline{a}\rangle\tau_{\overline{i}} \wedge [\overline{a}](\neg(\pi)_{\overline{i}} : a) \in s_l$. By Lemma 6.2(4) there exists $s_{l-1} \in \rho_s \cap W^{\overline{i}}$ such that $s_{l-1} \xrightarrow{a} s_l$. From Lemma 6.2(3), $\neg(\pi)_{\overline{i}} : a \in s_{l-1}$. Since s_{l-1} is an MCS, we have $(\pi)_{\overline{i}} : a \notin s_{l-1}$. By induction hypothesis, $a \notin \pi(s_{l-1})$, therefore we have $\rho^s_{s_0}$ does not conform to π.

($\Leftarrow$) Conversely, using (A6c) and a similar argument it can be shown that if $(\pi \Rightarrow \sigma')_i : c \notin s$ then $c \notin (\pi \Rightarrow \sigma')(s)$. Q.E.D.

Theorem 6.5. For all $\alpha \in \Pi$, for all $s \in S$, $\alpha \in s$ iff $M, s \models \alpha$.

Proof. The proof is by induction on the structure of α. $\alpha \equiv (\sigma)_i : c$. From Lemma 6.4 we have $(\sigma)_i : c \in s$ iff $c \in \sigma(s)$ iff by semantics $M, s \models (\sigma)_i : c$.

$\alpha \equiv \sigma \rightsquigarrow_i \beta$.

($\Rightarrow$) We show the following:

(1) If $\sigma \rightsquigarrow_i \beta \in s$ and there exits a transition $s \xrightarrow{a} s'$ such that $a \in \sigma(s)$, then $\{\beta, \sigma \rightsquigarrow_i \beta\} \subseteq s'$. Suppose $\sigma \rightsquigarrow_i \beta \in s$, from (A7) we have $\beta \in s$. We have two cases to consider.

- $s \in W^i$: We have $\tau_i \in s$. Since $a \in \sigma(s)$, by Lemma 6.4 we have $(\sigma)_i : a \in s$. From (A7) we get $[a](\sigma \rightsquigarrow_i \beta) \in s$. By Lemma 6.2(1) we have $\sigma \rightsquigarrow_i \beta \in s'$.
- $s \in W^{\overline{i}}$: We have $\tau_{\overline{i}} \in s$. From (A7) we get $[N](\sigma \rightsquigarrow_i \beta) \in s$, since s is an MCS we have for every $a \in \Sigma$, $[a](\sigma \rightsquigarrow_i \beta) \in s$. By Lemma 6.2(1) we have $\sigma \rightsquigarrow_i \beta \in s'$.

By applying (A7) at s' we get $\beta \in s'$.

(2) If $\sigma \rightsquigarrow_i \beta \in s$ and s is a non-leaf node, then $\exists s'$ such that $s \xrightarrow{a} s'$ and $a \in \sigma(s)$.

Suppose s is a non-leaf node. From (A7), $\bigvee_{a\in\Sigma}(\langle a\rangle\mathsf{True} \wedge (\sigma)_i : a) \in s$. Since s is an MCS, there exists an a such that $\langle a\rangle\mathsf{True} \wedge (\sigma)_i : a \in s$.

By Lemma 6.2(2), there exists an s' such that $s \stackrel{a}{\Longrightarrow} s'$ and by Lemma 6.4 $a \in \sigma(s)$.

(1) ensures that whenever $\sigma \rightsquigarrow_i \beta \in s$ and there exists a path $\rho_s^{s_k}$ which conforms to σ, then we have $\{\beta, \sigma \rightsquigarrow_i \beta\} \subseteq s_k$. Since $\beta \in \mathsf{Past}(P)$, by Lemma 6.3 we have $M, s_k \models \beta$. (2) ensures that for all paths $\rho_s^{s_k}$ which conforms to σ, if s_k is a non-leaf node, then $\mathsf{moves}(s) \cap \sigma(s) \neq \varphi$. Therefore we get $M, s \models \sigma \rightsquigarrow_i \beta$.

($\Leftarrow$) Conversely suppose $\sigma \rightsquigarrow_i \beta \notin s$, to show $M, s \not\models \sigma \rightsquigarrow_i \beta$. Suffices to show that there exists a path $\rho_s^{s_k}$ that conforms to σ such that $M, s_k \not\models \beta$ or s_k is a non-leaf node and $\mathsf{moves}(s_k) \cap \sigma(s_k) = \varphi$.

Lemma 6.6. For all $t \in \mathcal{AT}$, $\sigma \rightsquigarrow_i \beta \notin t$ implies there exists a path $\rho_t^{t_k} : t = t_1 \xrightarrow{a_1}_{\mathcal{AT}} t_2 \ldots \xrightarrow{a_{k-1}}_{\mathcal{AT}} t_k$ which conforms to σ such that one of the following conditions hold.

- $\beta \notin t_k$.
- t_k is a non-leaf node and $\mathsf{moves}(t_k) \cap \sigma(t_k) = \varphi$.

We have $t = s \cap \mathsf{CL}(\sigma \rightsquigarrow_i \beta)$ is an atom. By Lemma 6.6 (proof given in the Appendix), there exists a path in the atom graph $t = t_1 \xrightarrow{a_1}_{\mathcal{AT}} t_2 \ldots \xrightarrow{a_k}_{\mathcal{AT}} t_k$ such that $\beta \notin t_k$ or t_k is a non-leaf node and $\mathsf{moves}(t_k) \cap \sigma(t_k) = \varphi$. t_1 can be extended to the MCS s. Let $t_2' = t_2 \cup \{\alpha | [a_1]\alpha \in s\}$. Its easy to check that t_2' is consistent. Consider any MCS s_2 extending t_2', we have $s \stackrel{a_1}{\Longrightarrow} s_2$. Continuing in this manner we get a path in $s = s_1 \stackrel{a_1}{\Longrightarrow} s_2 \ldots \stackrel{a_{k-1}}{\Longrightarrow} s_k$ in M which conforms to σ where either $\beta \notin s_k$ or s_k is a non-leaf node and $\mathsf{moves}(s_k) \cap \sigma(s) = \varphi$. Q.E.D.

7 Extensions for Strategy Specification

Until operator

One of the natural extensions to strategy specification is to come up with a construct which asserts that a player strategy conforms to some specification σ until a certain condition holds. Once the condition is fulfilled, he is free to choose any action.

We can add the future modality $\Diamond\alpha$ in the logic defined in Section 3 with the following interpretation.

- $M, s \models \Diamond\gamma$ iff there exists an s' such that $s \Longrightarrow^* s'$ and $M, s' \models \gamma$.

Let $\mathsf{Past}(\Pi)$ and $\mathsf{Future}(\Pi)$ denote the past and future fragment of Π respectively. I.e.,

$$\mathsf{Past}(\Pi^{P^i}) := p \in P^i \mid \neg\alpha \mid \alpha_1 \vee \alpha_2 \mid \Diamond\!\!\!-\alpha$$

$$\mathsf{Future}(\Pi^{P^i}) := p \in P^i \mid \neg\alpha \mid \alpha_1 \vee \alpha_2 \mid \Diamond\alpha$$

Let $\Box\alpha = \neg\Diamond\neg\alpha$ and $\boxminus\alpha = \neg\diamondminus\neg\alpha$. We can enrich $\mathsf{Strat}^i(P^i)$ with the until operator $\sigma\mathbf{U}\varphi$, where $\varphi \in \mathsf{Past}(\Pi^{P^i}) \cup \mathsf{Future}(\Pi^{P^i})$, with the following interpretation:

- $(\sigma\mathbf{U}\varphi)(s) = \begin{cases} \Sigma & \text{if } \exists j : 0 \leq j \leq m \text{ such that } \rho_{s_0}^{s_j}, j \models \varphi \\ \sigma(s) & \text{otherwise} \end{cases}$

Note that until does not guarantee that φ will eventually hold. We can extend the axiomatization quite easily to handle the new construct. Firstly we need to add the following axiom and the derivation rule for the future modality.

$$\Box\alpha \equiv (\alpha \wedge [N]\Box\alpha) \qquad \text{(Ax-box)}$$

$$\frac{\alpha \supset [N]\alpha}{\alpha \supset \Box\alpha} \qquad \text{(Ind)}$$

Using the above axiom and inference rule one can easily show the analogue of Lemma 6.2 and Lemma 6.3 for the future modality. For the until operator we have the following axiom.

$$(\sigma\mathbf{U}\varphi)_i : c \equiv \neg\diamondminus\varphi \supset (\sigma)_i : c \qquad \text{(Ax-Until)}$$

We can show that Lemma 6.4 holds once again, for the extended syntax:

Lemma 7.1. For all i, for all $\sigma \in \mathsf{Strat}^i(P^i)$, for all $c \in \Sigma$, for all $s \in S$, $(\sigma)_i : c \in s$ iff $c \in \sigma(s)$.

Proof. The proof is by induction on the structure of σ as seen before. The interesting case is when $\sigma \equiv \sigma'\mathbf{U}\varphi$:

($\Rightarrow$) Suppose $(\sigma'\mathbf{U}\varphi)_i : c \in s$. It suffices to show that $\forall j : 0 \leq j \leq k$, $\rho_{s_0}^{s_j}, j \not\models \varphi$ implies $c \in \sigma'(s)$. From axiom (Ax-Until), we have $\neg\diamondminus\varphi \supset (\sigma')_i : c \in s$. Rewriting this, we get $\diamondminus\varphi \in s$ or $(\sigma')_i : c \in s$.

- if $\diamondminus\varphi \in s$, then by Lemma 6.2, $\exists j : 0 \leq j \leq k$ such that $\varphi \in s_j$. Therefore we have $\rho_{s_0}^{s_j} \models \varphi$.
- if $(\sigma)_i : c \in s$, then by induction hypothesis we have $c \in \sigma(s)$.

($\Leftarrow$) To show $(\sigma'\mathbf{U}\varphi)_i : c \notin s$ implies $c \notin (\sigma'\mathbf{U}\varphi)(s)$. It suffices to show that $\forall j : 0 \leq j \leq m$, $\rho_{s_0}^{s_j}, j \not\models \varphi$ and $c \notin \sigma'(s)$. From axiom (Ax-Until), we have $\neg\diamondminus\varphi \wedge \neg((\sigma')_i : c) \in s$. Rewriting this we get $\boxminus\neg\varphi \in s$ and $\neg((\sigma)_i : c) \in s$.

- $\boxminus\neg\varphi \in s$ implies $\forall j : 0 \leq j \leq m$, $\neg\varphi \in s_j$(by Lemma 6.2). Since s_j is an MCS, $\alpha \notin s_j$. Therefore we have $\forall j : 0 \leq j \leq m$, $\rho_{s_0}^{s_j}, j \not\models \varphi$.
- $\neg((\sigma)_i : c) \in s$ implies $(\sigma)_i : c \notin s$(Since s is an MCS). By induction hypothesis we have $c \notin \sigma(s)$.

Q.E.D.

Nested strategy specification

Instead of considering simple past time formulas as conditions to be verified before deciding on a move, we can enrich the structure to assert the opponents conformance to some strategy specification in the history of the play. This can be achieved by allowing nesting of strategy specification. We can extend the strategy specification syntax to include nesting as follows.

$$\Gamma^i := \psi \mid \sigma \mid \gamma_1 \wedge \gamma_2$$
$$\mathsf{Strat}^i_{\mathrm{rec}}(P^i) := [\gamma \mapsto a]^i \mid \sigma_1 + \sigma_2 \mid \sigma_1 \cdot \sigma_2$$

where $\psi \in \mathsf{Past}(P^i)$, $\sigma \in \mathsf{Strat}^i_{\mathrm{rec}}(P^i)$ and $\gamma \in \Gamma^{\overline{i}}$. Below we give the semantics for the part that requires change. For the game tree $\mathcal{T}$ and a node s in it, let $\rho^s_{s_0} : s_0 \stackrel{a_1}{\Longrightarrow} s_1 \cdots \stackrel{a_m}{\Longrightarrow} s_m = s$ denote the unique path from s_0 to s

- $[\gamma \mapsto a]^i(s) = \begin{cases} a & \text{if } \ s \in W^i \text{ and } \rho^s_{s_0}, m \models \gamma \\ \Sigma & \text{otherwise} \end{cases}$
- $\rho^s_{s_0}, m \models \sigma$ iff $\forall j : 0 \leq j < m$, $a_i \in \sigma(s_i)$.
- $\rho^s_{s_0}, m \models \gamma_1 \wedge \gamma_2$ iff $\rho^s_{s_0}, m \models \gamma_1$ and $\rho^s_{s_0}, m \models \gamma_2$.

For a past formula ψ, the notion of $\rho^s_{s_0}, m \models \psi$ is already defined in Section 3. Let L denote the logic introduced in Section 3 and L_{rec} be the same as L except that $\sigma \in \mathit{Strat}^i_{\mathrm{rec}}(P^i)$. We show that L and L_{rec} have equivalent expressive power. Therefore one can stick to the relatively simple strategy specification syntax given in Section 3 rather than taking into account explicit nesting.

It is easy to see that any formula $\gamma \in \Gamma^i$, can be rewritten in the form $\sigma' \wedge \psi$ where $\sigma' \in \mathsf{Strat}^i_{\mathrm{rec}}(P^i)$ and $\psi \in \mathsf{Past}(P^i)$. This is due to the fact that if $\psi_1, \psi_2 \in \mathsf{Past}(P^i)$ then $\psi_1 \wedge \psi_2 \in \mathsf{Past}(P^i)$ and $\sigma_1 \wedge \sigma_2 \equiv \sigma_1 \cdot \sigma_2$(formally $\forall s, \rho^s_{s_0}, m \models \sigma_1 \wedge \sigma_2$ iff $\rho^s_{s_0}, m \models \sigma_1 \cdot \sigma_2$).

Given $\sigma_{\mathrm{rec}} \in \mathsf{Strat}^i_{\mathrm{rec}}(P^i)$ the equivalent formula $\sigma \in \mathit{Strat}^i(P^i)$ is constructed inductively as follows.

$$[\![[\psi \mapsto a]]\!] = [\psi \mapsto a]$$
$$[\![\sigma_1 + \sigma_2]\!] = [\![\sigma_1]\!] + [\![\sigma_2]\!]$$
$$[\![\sigma_1 \cdot \sigma_2]\!] = [\![\sigma_1]\!] \cdot [\![\sigma_2]\!]$$
$$[\![[\sigma \mapsto a]]\!] = [\![\sigma]\!] \Rightarrow [\mathsf{True} \mapsto a]$$
$$[\![[\sigma \wedge \psi \mapsto a]]\!] = [\![\sigma]\!] \Rightarrow [\psi \mapsto a]$$

Lemma 7.2. For all i, for all $s \in S$, for all $\sigma \in \mathit{Strat}^i(P^i)$, $\sigma(s) = [\![\sigma]\!](s)$.

Proof. The proof is by induction on the structure of σ. Let $s \in S$ and $\rho^s_{s_0} : s_0 \stackrel{a_1}{\Longrightarrow} s_1 \cdots \stackrel{a_m}{\Longrightarrow} s_m = s$ be the unique path from root to s.

$\sigma \equiv [\psi \mapsto a]$: Follows from the definition.

$\sigma \equiv \sigma_1 \cdot \sigma_2$ and $\sigma \equiv \sigma_1 + \sigma_2$ follows easily by applying induction hypothesis.

$\sigma \equiv [\pi \mapsto a]$: We need to show that for all s, $[\pi \mapsto a](s) = (\llbracket\pi\rrbracket \Rightarrow [\mathsf{True} \mapsto a])(s)$. We have the following two cases:

- $\rho^s_{s_0}, m \models \pi$: In this case, we have $[\pi \mapsto a](s) = a$. $\rho^s_{s_0}, m \models \pi$ implies $\forall j : 0 \leq j < m$, $a_j \in \pi(s_j)$. From induction hypothesis, $a_j \in \llbracket\pi\rrbracket(s_j)$, which implies $\rho^s_{s_0}$ conforms to $\llbracket\pi\rrbracket$. From the semantics, we get $(\llbracket\pi\rrbracket \Rightarrow [\mathsf{True} \mapsto a])(s) = ([\mathsf{True} \mapsto a])(s) = a$.

- $\rho^s_{s_0}, m \not\models \pi$: In this case, we have $[\pi \mapsto a](s) = \Sigma$ and $\exists j : 0 \leq j < m$ such that $a_j \notin \pi(s_j)$. By induction hypothesis, we have $a_j \notin \llbracket\pi\rrbracket(s_j)$ which implies that $\rho^s_{s_0}$ does not conform to $\llbracket\pi\rrbracket$. From semantics we get that $(\llbracket\pi\rrbracket \Rightarrow [\mathsf{True} \mapsto a])(s) = \Sigma$.

$\sigma \equiv [\pi \wedge \psi \mapsto a]$: The following two cases arise:

- $\rho^s_{s_0}, m \models \pi \wedge \psi$: We have $[\pi \wedge \psi \mapsto a](s) = a$. $\rho^s_{s_0}, m \models \pi \wedge \psi$ implies $\rho^s_{s_0}, m \models \pi$ and $\rho^s_{s_0}, m \models \psi$. $\rho^s_{s_0}, m \models \pi$ implies $\forall j : 0 \leq j < m$, $a_j \in \pi(s_j)$. By induction hypothesis, $a_j \in \llbracket\pi\rrbracket(s_j)$ and as before we get $(\llbracket\pi\rrbracket \Rightarrow [\psi \mapsto a])(s) = ([\psi \mapsto a])(s) = a$.

- $\rho^s_{s_0}, m \not\models \pi \wedge \psi$: We have the following two cases:

  - $\rho^s_{s_0}, m \not\models \psi$: It is easy to see that

    $$\pi \wedge \psi \mapsto a](s) = (\llbracket\pi\rrbracket \Rightarrow [\psi \mapsto a])(s) = \Sigma.$$

  - $\rho^s_{s_0}, m \not\models \pi$: In this case, $\exists j : 0 \leq j < m$ such that $a_j \notin \pi(s_j)$. By induction hypothesis, we have $a_j \notin \llbracket\pi\rrbracket(s_j)$. By an argument similar to the one above we get $\llbracket\pi\rrbracket \Rightarrow [\psi \mapsto a])(s) = \Sigma$.

Q.E.D.

For the converse, given a $\sigma \in \mathsf{Strat}^i(P^i)$, we can construct an equivalent formula $\sigma_{\mathrm{rec}} \in \mathsf{Strat}^i_{\mathrm{rec}}(P^i)$. The crucial observation is the following equivalences in $\mathsf{Strat}^i(P^i)$.

- $\pi \Rightarrow \sigma_1 + \sigma_2 \equiv (\pi \Rightarrow \sigma_1) + (\pi \Rightarrow \sigma_2)$

- $\pi \Rightarrow \sigma_1 \cdot \sigma_2 \equiv (\pi \Rightarrow \sigma_1) \cdot (\pi \Rightarrow \sigma_2)$

- $\pi_1 \Rightarrow (\pi_2 \Rightarrow \sigma) \equiv (\pi_1 \cdot \pi_2) \Rightarrow \sigma$

Using the above equivalences, we can write the strategy specification σ in a normal form where all the implications are of the form $\pi \Rightarrow [\psi \mapsto a]$. Then σ_{rec} is constructed inductively as follows:

$$\begin{aligned} [\![[\psi \mapsto a]]\!] &= [\psi \mapsto a] \\ [\![\sigma_1 + \sigma_2]\!] &= [\![\sigma_1]\!] + [\![\sigma_2]\!] \\ [\![\sigma_1 \cdot \sigma_2]\!] &= [\![\sigma_1]\!] \cdot [\![\sigma_2]\!] \\ [\![\pi \Rightarrow [\psi \mapsto a]]\!] &= [[\![\pi]\!] \wedge \psi \mapsto a] \end{aligned}$$

Lemma 7.3. For all i, for all $s \in S$, for all $\sigma \in Strat^i(P^i)$, $\sigma(s) = [\![\sigma]\!](s)$.

Proof. The proof is by induction on the structure of formula. Let $\rho^s_{s_0} : s_0 \stackrel{a_1}{\Longrightarrow} s_1 \cdots \stackrel{a_m}{\Longrightarrow} s_m = s$. The interesting case is when $\sigma \equiv \pi \Rightarrow [\psi \mapsto a]$. We need to show that for all s, $\pi \Rightarrow [\psi \mapsto a](s) = [\pi \wedge \psi \mapsto a](s)$. We have the following two cases:

- $\rho^s_{s_0}$ conform to π: We have $\pi \Rightarrow [\psi \mapsto a](s) = [\psi \mapsto a](s)$ and $\forall j : 0 \leq j < m$, $a_j \in \pi(s_i)$. By induction hypothesis, $a_j \in [\![\pi]\!](s_i)$ which implies that $\rho^s_{s_0}, m \models \pi$. Therefore $[[\![\pi]\!] \wedge \psi \mapsto a](s) = [\psi \mapsto a](s)$.
- $\rho^s_{s_0}$ does not conform to π: By an argument similar to the above, we can show that $\pi \Rightarrow [\psi \mapsto a](s) = [\pi \wedge \psi \mapsto a](s) = \Sigma$.

Q.E.D.

Theorem 7.4. Logics L and L_{rec} have equivalent expressive power. I.e.,

- For every $\alpha \in \Pi$, there exists $\alpha_{\mathrm{rec}} \in \Pi_{\mathrm{rec}}$ such that $M, s \models \alpha$ iff $M, s \models \alpha_{\mathrm{rec}}$.
- For every $\alpha_{\mathrm{rec}} \in \Pi_{\mathrm{rec}}$ there exists $\alpha \in \Pi$ such that $M, s \models \alpha_{\mathrm{rec}}$ iff $M, s \models \alpha$.

Proof. The theorem follows from Lemma 7.2 and Lemma 7.3 by a routine inductive argument. Q.E.D.

8 Discussion

We have defined a logic for reasoning about composite strategies in games. We have presented an axiomatization for the logic and shown its completeness.

We again remark that the presentation has been given for two-player games only for easy readability. It can be checked that all the definitions and arguments given here can be appropriately generalized for n-player games.

While our emphasis in the paper has been on advocating syntactically constructed strategies, we make no claims to having the "right" set of connectives for building them. This will have to be decided by experience, gained by specifying several kinds of strategies which turn out to be of use in reasoning about games.

We believe that a framework of this sort will prove useful in reasoning about multi-stage and *repeated* games, where strategy revision based on learning other players' strategies (perhaps partially) plays an important role.

Appendix

Lemma A.1. For atoms t_1 and t_2, the following statements are equivalent.

1. $\widehat{t_1} \wedge \langle a \rangle \widehat{t_2}$ is consistent.
2. $\langle \overline{a} \rangle \widehat{t_1} \wedge \widehat{t_2}$ is consistent.

Proof. Suppose $\langle \overline{a} \rangle \widehat{t_1} \wedge \widehat{t_2}$ is consistent, from (A3b) we have $\langle \overline{a} \rangle \widehat{t_1} \wedge [\overline{a}] \langle a \rangle \widehat{t_2}$ is consistent. Therefore, $\langle \overline{a} \rangle (\widehat{t_1} \wedge \langle a \rangle \widehat{t_2})$ is consistent, which implies $\not\vdash [\overline{a}] \neg (\widehat{t_1} \wedge \langle a \rangle \widehat{t_2})$. From (NG-), $\not\vdash \neg (\widehat{t_1} \wedge \langle a \rangle \widehat{t_2})$, thus we have that $\widehat{t_1} \wedge \langle a \rangle \widehat{t_2}$ is consistent.

Suppose $\widehat{t_1} \wedge \langle a \rangle \widehat{t_2}$ is consistent, from (A3a) we have $[a] \langle \overline{a} \rangle \widehat{t_1} \wedge \langle a \rangle \widehat{t_2}$ is consistent. Therefore, $\langle a \rangle (\langle \overline{a} \rangle \widehat{t_1} \wedge \widehat{t_2})$ is consistent, which implies $\not\vdash [a] \neg (\langle \overline{a} \rangle \widehat{t_1} \wedge \widehat{t_2})$. From (NG-), $\not\vdash \neg (\langle \overline{a} \rangle \widehat{t_1} \wedge \widehat{t_2})$, thus we get that $\langle \overline{a} \rangle \widehat{t_1} \wedge \widehat{t_2}$ is consistent. Q.E.D.

Lemma A.2. Consider the path $t_k \xrightarrow{a_k} t_{k-1} \ldots \xrightarrow{a_1} t_0$ where t_k is a root atom.

1. For all $j \in \{0, \ldots, k-1\}$, if $[\overline{a}]\alpha \in t_j$ and $t_{j+1} \xrightarrow{a} t_j$ then $\alpha \in t_{j+1}$.
2. For all $j \in \{0, \ldots, k-1\}$, if $\langle \overline{a} \rangle \alpha \in t_j$ and $t_{j+1} \xrightarrow{b} t_j$ then $b = a$ and $\alpha \in t_{j+1}$.
3. For all $j \in \{0, \ldots, k-1\}$, if $\diamond\!\!\!-\alpha \in t_j$ then there exists $i : j \leq i \leq k$ such that $\alpha \in t_i$.

Proof.

(1) Since $t_{j+1} \xrightarrow{a} t_j$, we have $\widehat{t_{j+1}} \wedge \langle a \rangle \widehat{t_j}$ is consistent, By Lemma A.1, $\widehat{t_j} \wedge \langle \overline{a} \rangle \widehat{t_{j+1}}$ is consistent, which implies $[\overline{a}]\alpha \wedge \langle \overline{a} \rangle \widehat{t_{j+1}}$ is consistent (by omitting some conjuncts). Therefore $\langle \overline{a} \rangle (\alpha \wedge \widehat{t_{j+1}})$ is consistent. Using (NG-) we get $\alpha \wedge \widehat{t_{j+1}}$ is consistent and since t_{j+1} is an atom, we have $\alpha \in t_{j+1}$.

(2) Suppose $t_{j+1} \xrightarrow{b} t_j$, we first show that $b = a$. Suppose this is not true, since $t_{j+1} \xrightarrow{b} t_j$, we have $\widehat{t_j} \wedge \langle \overline{b} \rangle \widehat{t_{j+1}}$ is consistent. And therefore $\widehat{t_j} \wedge \langle \overline{b} \rangle \mathsf{True}$

is consistent. From axiom (A2c) $\widehat{t_j} \wedge [\overline{a}]\mathsf{False}$ is consistent. If $\langle\overline{a}\rangle\alpha \in t_j$, then we get $\langle\overline{a}\rangle\alpha \wedge [\overline{a}]\mathsf{False}$ is consistent. Therefore $\langle\overline{a}\rangle(\alpha \wedge \mathsf{False})$ is consistent. From (NG-) we have $\alpha \wedge False$ is consistent, which is a contradiction.

To show $\alpha \in t_{j+1}$ observe that $\langle\overline{a}\rangle\alpha \in t_j$ implies $[\overline{a}]\alpha \in t_j$ (by axiom (A2b) and closure condition). By previous argument we get $\alpha \in t_{j+1}$.

(3) Suppose $\Diamond\alpha \in t_j$ and $t_{j+1} \xrightarrow{a} t_j$. If $\alpha \in t_j$ then we are done. Else by axiom (A4b) and the previous argument, we have $\langle\overline{a}\rangle\Diamond\alpha \in t_j$. From (2) we have $\Diamond\alpha \in t_{j+1}$. Continuing in this manner, we either get an i where $\alpha \in t_i$ or we get $\Diamond\alpha \in t_k$. Since t_k is the root, this will give us $\alpha \in t_k$. Q.E.D.

Lemma A.3. For all $t \in \mathcal{AT}$, $\sigma \rightsquigarrow_i \beta \notin t$ implies there exists a path $\rho_t^{t_k} : t = t_1 \xrightarrow{a_1}_{\mathcal{AT}} t_2 \ldots \xrightarrow{a_{k-1}}_{\mathcal{AT}} t_k$ which conforms to σ such that one of the following conditions holds:

- $\beta \notin t_k$.
- t_k is a non-leaf node and $\mathsf{moves}(t_k) \cap \sigma(t_k) = \varphi$.

Proof. Consider the least set R containing t and closed under the following condition:

- if $t_1 \in R$ then for every transition $t_1 \xrightarrow{a} t_2$ such that $a \in \sigma(t_1)$ we have $t_2 \in R$.

If there exists an atom $t' \in R$ such that $\beta \notin t'$ or if t' is a non-leaf node and $\mathsf{moves}(t') \cap \sigma(t') = \varphi$, then we are done. Suppose not, then we have $\vdash \widetilde{R} \supset \beta$ and $\vdash (\widetilde{R} \wedge \neg\mathsf{leaf}) \supset \bigvee_{a\in\Sigma}(\langle a\rangle\mathsf{True} \wedge (\sigma)_i : a)$.

Claim A.4. The following are derivable.

1. $\vdash (\widetilde{R} \wedge \tau_i \wedge (\sigma)_i : a) \supset [a]\widetilde{R}$.
2. $\vdash (\tau_{\overline{i}} \wedge \widetilde{R}) \supset [N]\widetilde{R}$.

Assume claim A.4 holds, then applying (IND) rule we get $\vdash \widetilde{R} \supset \sigma \rightsquigarrow_i \beta$. But $t \in R$ and therefore $\vdash \widehat{t} \supset \sigma \rightsquigarrow_i \beta$, contradicting the assumption that $\sigma \rightsquigarrow_i \beta \notin t$. Q.E.D.

Proof. To prove 1, suppose the claim does not hold. We have that $(\widetilde{R} \wedge \tau_i \wedge (\sigma)_i : a) \wedge \langle a\rangle\neg\widetilde{R}$ is consistent. Let $R' = \mathcal{AT} - R$. If $R' = \varphi$ then $R = \mathcal{AT}$ in which case its easy to see that the claim holds. If $R' \neq \varphi$, then we have $(\widetilde{R} \wedge \tau_i \wedge (\sigma)_i : a) \wedge \langle a\rangle\widetilde{R'}$ is consistent. Hence for some $t_1 \in R$ and $t_2 \in R'$, we have $(\widehat{t_1} \wedge \tau_i \wedge (\sigma)_i : a)) \wedge \langle a\rangle\widehat{t_2}$ is consistent. Which implies $t_1 \xrightarrow{a}_{\mathcal{AT}} t_2$ and this transition conforms to σ. By closure condition on R, $t_2 \in R$ which gives us the required contradiction.

Proof of 2 is similar. Q.E.D. (Claim A.4)

References

[1] R. Alur, T.A. Henzinger & O. Kupferman. Alternating-time temporal logic. In W.P. de Roever, H. Langmaack & A. Pnueli, eds., *Compositionality: The Significant Difference, International Symposium, COMPOS'97, Bad Malente, Germany, September 8–12, 1997. Revised Lectures*, vol. 1536 of *Lecture Notes in Computer Science*, pp. 23–60. Springer, 1998.

[2] J. van Benthem. Games in dynamic epistemic logic. *Bulletin of Economic Research*, 53(4):219–248, 2001.

[3] J. van Benthem. Logic games are complete for game logics. *Studia Logica*, 75(2):183–203, 2003.

[4] J. van Benthem. An essay on sabotage and obstruction. In D. Hutter & W. Stephan, eds., *Mechanizing Mathematical Reasoning*, vol. 2605 of *Lecture Notes in Computer Science*, pp. 268–276. Springer, 2005.

[5] G. Bonanno. The logic of rational play in games of perfect information. *Economics and Philosophy*, 7:37–65, 1991.

[6] H.P. van Ditmarsch. *Knowledge Games*. Ph.D. thesis, University of Groningen, 2000. *ILLC Publications* DS-2000-06.

[7] A. Ehrenfeucht. An application of games to the completeness problem for formalized theories. *Fundamenta Mathematicae*, 49:129–141, 1961.

[8] V. Goranko. Coalition games and alternating temporal logics. In J. van Benthem, ed., *Proceedings of the 8th Conference on Theoretical Aspects of Rationality and Knowledge (TARK-2001), Certosa di Pontignano, University of Siena, Italy, July 8–10, 2001*, pp. 259–272. Morgan Kaufmann, 2001.

[9] D. Harel, D. Kozen & J. Tiuryn. *Dynamic Logic*. The MIT Press, October 2000.

[10] P. Harrenstein, W. van der Hoek, J.-J. Meyer & C. Witteven. A modal characterization of Nash equilibrium. *Fundamenta Informaticae*, 57(2–4):281–321, 2003.

[11] W. van der Hoek, W. Jamroga & M. Wooldridge. A logic for strategic reasoning. In F. Dignum, V. Dignum, S. Koenig, S. Kraus, M.P. Singh & M. Wooldridge, eds., *4rd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2005), July 25–29, 2005, Utrecht, The Netherlands*, pp. 157–164. ACM, 2005.

[12] M. Lange. *Games for modal and temporal logics*. Ph.D. thesis, University of Edinburgh, 2002.

[13] R. Parikh. The logic of games and its applications. *Annals of Discrete Mathematics*, 24:111–140, 1985.

[14] M. Pauly. *Logic for Social Software*. Ph.D. thesis, University of Amsterdam, October 2001. *ILLC Publications* DS-2001-10.

Models of Awareness

Giacomo Sillari

Philosophy, Politics and Economics Program
University of Pennsylvania
Philadelphia PA 19104, United States of America
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.

1 Introduction

Since its first formulations (cf. [21]), 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

Giacomo Bonanno, Wiebe van der Hoek, Michael Wooldridge (eds.). *Logic and the Foundations of Game and Decision Theory (LOFT 7)*. Texts in Logic and Games 3, Amsterdam

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 [9], Fagin and Halpern formally introduced the concept of awareness in the context of epistemic logic, providing semantic grounds for the distinction between implicit and explicit belief. The technical concept of "awareness" they introduce is amenable to different discursive interpretations that can 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. [32, 33]), and noticed that partitional structures as introduced by Aumann cannot represent awareness [8]. The model in [33] 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. Halpern ([13]) shows that such a model is a particular case of the logic of awareness introduced in [9]. Heifetz et al. ([20]) present a set-theoretical model that generalizes traditional information structures *à la* Aumann. Its axiomatization in a 3-valued epistemic logic is provided by by Halpern and Rêgo in [14]. A further, purely set-theoretical model of awareness is given by Li ([30]). 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. Feinberg ([11]) 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 is performed by Sillari ([37]). Haleprn and Rêgo ([15]) define extensive-form games with possibly unaware players in which the usual assumption of common knowledge of the structure of the game may fail. Heifetz et al. ([19]) take into account Bayesian games with unawareness. In [29] the concept of subgame perfect equilibrium is extended to games with unawareness.

This paper makes two main contributions 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. [25, pp. 9–13]). Furthermore, I consider Newell's theory of agency (as advanced in [35]) 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 [35, 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 [21] 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, [16] 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 [16] 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. [28] and the reconstruction offered in [37], in which awareness structures find a concrete application.

enumerable, even in the case of **S5**, which was proven to be decidable in [12]. 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, [1] 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 [40] and [39], 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. [6]), 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. [2], [3]) 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. [3]) 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. [2]), 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 Section 2, 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 the intuition underlying Fagin and Halpern's logic of awareness. In Section 3, 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. [40]) can be modified to carry over to quantified logic of awareness.

2 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.

2.1 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 in [21, 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 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.

[2] Cf. [21], 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 mentions a natural language expression 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.

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 two faces of the same coin. Hintikka's concept of *defensibility*, intended as "immunity from certain standards of criticism" ([21, 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 [9].

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 boolean connectives, as a set μ of formulas such that

$$p \in \mu \rightarrow \neg p \notin \mu \qquad (\neg)$$
$$(p \wedge q) \in \mu \rightarrow p \in \mu \text{ and } q \in \mu \qquad (\wedge)$$
$$(p \vee q) \in \mu \rightarrow p \in \mu \text{ or } q \in \mu \qquad (\vee)$$
$$\neg\neg p \in \mu \rightarrow p \in \mu \qquad (\neg\neg)$$
$$\neg(p \wedge q) \in \mu \rightarrow \neg p \in \mu \text{ or } \neg q \in \mu \qquad (\neg\wedge)$$
$$\neg(p \vee q) \rightarrow \neg p \in \mu \text{ and } \neg q \in \mu \qquad (\neg\vee)$$

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

$$\text{If } P_i p \in \mu, \text{ then there exists at least a } \mu^* \text{ such that } \mu^* \text{ is an alternative to } \mu \text{ for } i, \text{ and } p \in \mu^* \qquad (P)$$
$$\text{If } K_i p \in \mu, \text{ then, if } \mu^* \text{ is an alternative to } \mu \text{ for } i, \text{ then } K_i p \in \mu^* \qquad (\text{KK})$$
$$\text{If } K_i p \in \mu, \text{ then } p \in \mu \qquad (\text{K})$$

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 close to 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 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 that can be actually or at least conceptually isolated sufficiently from the rest", cf. [23, p. 22]. But cf. also [21, pp. 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" (emphasis added).

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-sustenance*. 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, 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 [9], 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.

2.2 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 central 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

[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 [21, 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 [38]. 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 [24]) 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.

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 is reproduced in [35]. 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 laws that determine the behavior of the system. 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 insofar as 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

[8] Cf. [35, p. 100].

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 "slogan equation[11]"

$$\text{Representation} = \text{Knowledge} + \text{Access}.$$

The interesting question is then: in which way does an agent extract rep-

[9] Cf. [35, p. 102]. Although Newell, in the following, refines it, his principle of rationality does not seem to be explicitly concerned with utility.
[10] Cf. [35, p. 107].
[11] Cf. [35, p. 114].

resentation 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.

2.3 Levi: Commitment and performance

Levi illustrates (in [25]) 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] 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 sit-*

[12] The idea is taken up again in [5], 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 to obtain a working approximation of it.

[13] Cf. [35, 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.

[15] Cf. [25, pp. 9–10].

uated 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 [27]. 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 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 ap-

[16] Cf. [25, p. 12].

[17] She is committed in the sense (see [27]) 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"

pears 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 [25], 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.

2.4 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, [24]). In fact, Hintikka's position on the issue has shifted over the years. Since the introduction of Rantala's "urn models" (cf. [36]), the author of *Knowledge and Belief* has endorsed the "impossible worlds" solution to the problem of logical omniscience (cf. [22]). 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 elaborates 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 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

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.

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. [18]) and by Cherniak (cf. [7]). 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 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.

[18] Cf, [9, 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.", [26, p. 168, n. 14].

3 First-Order Logic of Awareness

In this section, I extend Fagin and Halpern's logic of general awareness (cf. [9]) 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. [31] and [10]. For detailed treatments of the first-order epistemic systems, cf. [6] and the work of Arló-Costa and Pacuit ([2] and [3]).

3.1 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:

$$\text{All axioms from } \mathbf{S} \tag{S}$$

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

$$\text{From } \varphi \rightarrow \psi \text{ infer } \varphi \rightarrow \forall x\psi, \text{ where } x \text{ is not free in } \varphi. \tag{Gen}$$

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 [6] 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.

3.2 Adding awareness

Following [9], 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

$$X_i\varphi \leftrightarrow K_i\varphi \wedge A_i\varphi. \tag{A0}$$

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 $(M, w) \models K_i\varphi$

In propositional awareness systems of the kind introduced in [9], 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.

[20] For the terminology, see [6].

[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.

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., [9]):

$$A_i \neg\varphi \leftrightarrow A_i\varphi \tag{A1}$$

$$A_i(\varphi \wedge \psi) \leftrightarrow A_i\varphi \wedge A_i\psi \tag{A2}$$

$$A_i K_j \varphi \leftrightarrow A_i\varphi \tag{A3}$$

$$A_i A_j \varphi \leftrightarrow A_i\varphi \tag{A4}$$

$$A_i X_j \varphi \leftrightarrow A_i\varphi. \tag{A5}$$

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* σ, 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 $\sigma(x) \in D_i(x)$.

3.3 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 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.

[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.

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) $X_i\exists x\neg A_iP(x)$ to be satisfiable, while

(ii) $\exists xX_i\neg A_iP(x)$ should not be satisfiable.

Semantically, thus, if $P(x) \notin \mathcal{A}_i(w)$, then (against (ii)), $\neg A_iP(x) \notin \mathcal{A}_i(w)$. Yet the possibility that (i) $\exists x\neg A_iP(x) \in \mathcal{A}_i(w)$ 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):

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

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 $(M, w) \models_{\sigma'} A_i\neg A_iP(x)$. But that is impossible, since, for quantifier-free propositions, awareness is generated by atomic formulas. On the other hand, (i) entails that $\exists x\neg A_iP(x) \in \mathcal{A}_i(w)$, which remains satisfiable, since the weak condition (*) does not require that $\neg A_iP(x) \in \mathcal{A}_i(w)$.

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 ($\neg A_iP(x)$), hence $A_i\exists x\neg A_iP(x)$ holds. Now, had (*) been a biconditional, since $\exists x\neg A_iP(x) \in \mathcal{A}_i(w)$ holds, it would have been the case that $\neg A_iP[x/y] \in \mathcal{A}_i(w)$, that is $A_i\neg A_iP(y)$. 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.

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

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

What is the interplay between universal quantification and awareness? Consider, in the example above, the situation in which White is aware that

[23] Cf. [16], in which a similar requirement of weak existential closure is used.

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

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

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 [3]. 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.

3.4 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].

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

[24] The *mosaic* technique on which the proof of the existence of an effective criterion for the validity problem is based was introduced by Németi (cf. for example [34]). The proof on which this subsection is based can be found in [40]. The same technique is used in [39] to show that first-order common knowledge logics are complete. For a more compact textbook exposition of the proof, cf. [4].

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 3.1 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 3.2. Fourth, the use of neighborhood structures slightly modifies the argument of the left to right direction in Theorem 3.3. 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 $\mathrm{sub}_x\,\varphi = \{\psi[x/y] : \psi(y) \in \mathrm{sub}\,\varphi\}$, where $\mathrm{sub}\,\varphi$ is the closure under negation of the set of subformulas of φ. Define a *type* t for φ as any boolean saturated subset of $\mathrm{sub}_x\,\varphi$[25], i.e., such that, (i) for all $\psi \in \mathrm{sub}_x\,\varphi$, $\neg\psi \in t$ iff $\psi \notin t$; and (ii) for all $\psi \wedge \chi \in \mathrm{sub}_x\,\varphi$, $\psi \wedge \chi \in t$ iff ψ and χ belong to t[26]. Types t and t' are said to *agree on* $\mathrm{sub}_0\,\varphi$ (the set of subsentences of φ) if $t \cap \mathrm{sub}_0\,\varphi = t' \cap \mathrm{sub}_0\,\varphi$.

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 $\mathrm{sub}_0\,\varphi$. Consider now a first-order structure $\mathcal{D} = (D, P_0^{\mathcal{D}}, \ldots)$, let $a \in D$ and define $t^{\mathcal{D}}(a) = \{\psi \in \mathrm{sub}_x\,\varphi : \mathcal{D} \models \overline{\psi}[a]\}$, 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 $T = \{t^{\mathcal{D}}(a) : a \in D\}$. Notice that T is a realizable world candidate iff a formula α_T is satisfiable in a first-order structure, where α_T is

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

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

[25] Or, equivalently, as "any subset of $\mathrm{sub}_x\,\varphi$ such that $\{\overline{\psi} : \psi \in t\}$ is maximal consistent", where ψ is any subformula of φ: cf. [4].

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

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 $K_i\psi \in \mathrm{sub}_x\,\varphi$, then, $K_i\psi \in r(w)$ iff $\{v : \psi \in r(v)\} \in \mathcal{N}(w)$,

(iii) if $A_i\psi \in \mathrm{sub}_x\,\varphi$, then $A_i\psi \in r(w)$ iff $\psi \in \mathcal{A}_i(w)$,

(iv) if $X_i\psi \in \mathrm{sub}_x\,\varphi$, 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 3.1. 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 A.1. 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].

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

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 A.2 of the Appendix, drawing from [17], describes how a monotonic neighborhood model can be unravelled in a tree-like model. Hence,

Lemma 3.2. 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 A.2. 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 $K_i\psi \in \mathrm{sub}_x\,\varphi$, then, $K_i\psi \in r(w_0)$ iff $\{v : \psi \in r(v)\} \in \mathcal{N}(w)$,

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

(iv) if $X_i\psi \in \mathrm{sub}_x\,\varphi$, 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 3.3. A monodic sentence φ is satisfiable iff there exists a satisfying set for φ, whose blocks contain a finite number of elements.

[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. Section 1), the restriction is not problematic.

Proof. See Appendix, Section A.3. Q.E.D.

The effective satisfiability criterion now follows:

Corollary 3.4. 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. [40]).

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 LOFT7 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.

Appendix

A.1 Proof of Theorem 3.1

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

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

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 $a \in D$, $t_a^v = \{\psi \in \mathrm{sub}_x\,\varphi : (M, v) \models_\sigma \psi(x)\}$ 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).

[$\Leftarrow$] 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] $r(w) = \{\psi \in \mathrm{sub}_x\, \varphi : I(w) \models \overline{\psi}[\langle r, \xi\rangle]\}$ 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 $\psi \in \mathrm{sub}\,\varphi$ 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

$I(w) \models_\sigma P_{K_i\chi}(y)$	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.

A.2 Unravelling a neighborhood structure

In this subsection I describe the procedure defined in [17] 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 A.1. (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 $C \subseteq X$.

[29] For a proof that this assumption is legitimate, cf. [39].

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 [17].

Define the following objects:

Definition A.2. 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:

$$S_0(X) = X, \qquad \mathcal{N}^c_0(X) = \bigcup_{x \in X} \mathcal{N}^c(x)$$

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

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 A.3. 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

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

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:

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

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

(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 $Y \subseteq W_{w_0}$:

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

Thus, every neighborhood in the original model $\mathcal{N}(\text{last}(\overrightarrow{s}))$ 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 $\varphi \in \mathcal{A}(\text{last}(\overrightarrow{w}))$.

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

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.

A.3 Proof of Theorem 3.3

Proof. [⇒] 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 $K\psi \in \mathrm{sub}_x\,\varphi$ 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 run such that $r(u) = t$. Consider the type $r(w)$ and the set $\mathcal{C} = \{\chi := K\psi \in \mathrm{sub}_x\,\varphi : \chi \notin r(w)\}$. 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 $2 \cdot |\,\mathrm{sub}_x\,\varphi| \cdot 2^{|\,\mathrm{sub}_x\,\varphi|}$ 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

[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 [40], 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.

the sequence thus constructed by defining the elements in $(W, \mathcal{N}, \mathcal{A}, f)$ as

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

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.

References

[1] A. Antonelli & R. Thomason. Representability in second-order polymodal logic. *Journal of Symbolic Logic*, 67(3):1039–1054, 2002.

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

[3] H. Arló-Costa & E. Pacuit. First-order classical model logic. *Studia Logica*, 84(2):171–210, 2006.

[4] T. Braüner & S. Ghlardi. First-order modal logic. In P. Blackburn, J. van Benthem & F. Wolter, eds., *Handbook of Modal Logic*, pp. 549–620. Elsevier, 2007.

[5] K. Carley & A. Newell. The nature of the social agent. *Journal of Mathematical Sociology*, 19(4):221–262, 1994.

[6] B. Chellas. *Modal Logic: An Introduction*. Cambridge University Press, Cambridge, 1980.

[7] C. Cherniak. *Minimal Rationality*. MIT Press, Cambridge, MA, 1986.

[8] E. Dekel, B.L. Lipman & A. Rustichini. Standard state-space models preclude unawareness. *Econometrica*, 66(1):159–173, 1999.

[9] R. Fagin & J. Halpern. Belief, awareness, and limited reasoning. *Artificial Intelligence*, 34(1):39–76, 1988.

[10] R. Fagin, J. Halpern, Y. Moses & M. Vardi. *Reasoning about Knowledge*. The MIT Press, Cambridge, MA, 1995.

[11] Y. Feinberg. Games with incomplete awareness. Tech. rep., Stanford University, 2005. Stanford Graduate School of Business Research Paper No. 1894.

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

[13] J. Halpern. Alternative semantics for unawareness. *Games and Economic Behavior*, 37(2):321–339, 2001.

[14] J. Halpern & L. Rêgo. Interactive unawareness revisited. In R. van der Meyden, ed., *Proceedings of the 10th Conference on Theoretical Aspects of Rationality and Knowledge (TARK-2005), Singapore, June 10–12, 2005*, pp. 78–91. National University of Singapore, 2005.

[15] J. Halpern & L. Rêgo. Extensive games with possibly unaware players. In H. Nakashima, M.P. Wellman, G. Weiss & P. Stone, eds., *5th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2006), Hakodate, Japan, May 8–12, 2006*, pp. 744–751. ACM, 2006.

[16] J. Halpern & L. Rêgo. Reasoning about knowledge of unawareness. In P. Doherty, J. Mylopoulos & C.A. Welty, eds., *Proceedings, Tenth International Conference on Principles of Knowledge Representation and Reasoning, Lake District of the United Kingdom, June 2–5, 2006*. AAAI Press, 2006.

[17] H.H. Hansen. *Monotonic Modal Logics*. Master's thesis, University of Amsterdam, 2003. *ILLC Publications* PP-2003-26.

[18] G. Harman. *Change in View*. MIT Press, Cambridge, MA, 1986.

[19] A. Heifetz, M. Meier & B. Schipper. Unawareness, beliefs and games. In D. Samet, ed., *Proceedings of the 11th Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pp. 183–192. UCL Presses Universitaires de Louvain, 2007.

[20] A. Heifetz, M. Meier & B.C. Schipper. Interactive unawareness. *Journal of Economic Theory*, 130(1):78–94, 2006.

[21] J. Hintikka. *Knowledge and Belief*. Cornell University Press, Ithaca, NY, 1962.

[22] J. Hintikka. Impossible possible worlds vindicated. *Journal of Philosophical Logic*, 4(3):475–484, 1975.

[23] J. Hintikka. Intellectual autobiography. In R.E. Auxier & L.E. Hahn, eds., *The Philosophy of Jaakko Hintikka*, vol. XXX of *The Library of Living Philosophers*, pp. 3–84. Open Court, 2003.

[24] H.-J. Levesque. A logic for implicit and explicit belief. In R.J. Brachman, ed., *Proceedings of the National Conference on Artificial Intelligence (AAAI-84), Austin, TX, August 6–10, 1984*, pp. 198–202. AAAI Press, 1984.

[25] I. Levi. *The Enterprise of Knowledge*. The MIT Press, Cambridge, MA, 1980.

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

[27] I. Levi. *The Covenant of Reason*. Cambridge University Press, Cambridge, 1997.

[28] D. Lewis. *Convention: A Philosophical Study*. Harvard University Press, Cambridge, MA, 1969.

[29] J. Li. Dynamic games of complete information with unawareness, 2006. Manuscript.

[30] J. Li. Information structures with unawareness, 2006. Manuscript.

[31] J.-J.Ch. Meyer & W. van der Hoek. *Epistemic Logic for AI and Computer Science*. Cambridge University Press, Cambridge, MA, 1995.

[32] S. Modica & A. Rustichini. Awareness and partitional information structures. *Theory and Decision*, 37(1):107–124, 1994.

[33] S. Modica & A. Rustichini. Unawareness and partitional information structures. *Games and Economic Behavior*, 27(2):265–298, 1999.

[34] I. Nemeti. Decidability of weakened versions of first-order logic. In L. Csirmaz, D.M. Gabbay & M. de Rijke, eds., *Logic Colloquim '92*, no. 1 in Studies in Logic, Language and Information, pp. 177–242. CSLI Publications, 1995.

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

[36] V. Rantala. Urn models: a new kind of non-standard model for first-order logic. *Journal of Philosophical Logic*, 4(3):455–474, 1975.

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

[38] G. Sillari. Quantified logic of awareness and impossible possible worlds, 2007. Manuscript.

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

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