'One is a Lonely Number': on the logic of communication

Johan van Benthem, Amsterdam & Stanford, December 2002

Logic Colloquium 2002

Abstract Logic is not just about single-agent notions like reasoning, or zero-agent notions like truth, but also about communication between two or more people. What we tell and ask each other can be just as 'logical' as what we infer in Olympic solitude. We show how such interactive phenomena can be studied systematically by merging epistemic and dynamic logic.

1 Logic in a social setting

1.1 Questions and answers

Consider the simplest type of communication: a question–answer episode between two agents. Here is a typical example. Being a Batavian soldier – a German tribe in the Rhine delta of proverbial valour – I approach you in a busy Roman street, A.D. 160, intent on contacting my revered general Maximus, now a captive, and ask:

Q *Is this the road to the Colosseum?*

As a well-informed and helpful Roman citizen, you answer

A *Yes.*

This is the sort of thing that we all do competently millions of times in our lives. There is nothing to it. But what is going on? I learn the fact that this is the road to the Colosseum. But much more happens. By asking the question, I convey to you that I do not know the answer, and also, that I think it possible that you do know. This information flows before you have said anything at all. Then, by answering, you do not just convey the topographical fact to me. You also bring it about that you know that I know, I know that you know I know, etc. This knowledge up to every finite depth of mutual reflection is called *common knowledge*. It involves a mixture of factual information and iterated information about what others know.

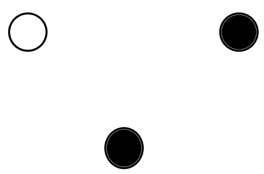
These *epistemic overtones* concerning our mutual information are not mere side-effects. They may steer further concrete actions. Some bystanders' knowing that I know may lead them to rush off and warn the Emperor Commodus – my knowing that they know I know may lead me to prevent them from doing just that. So epistemic overtones are ubiquitous and important, and we are good at computing them! In particular, we are well-attuned to fine differences in group knowledge. Everyone's knowing individually that your partner is unfaithful is unpleasant, but shame explodes when you meet people and know they all know that they know.

This is just the tip of an iceberg. I have described one type of question, but there are others. If you are my student, you would not assume that my classroom question shows that I do not know the answer. It need not even convey that I think you know, since my purpose may be to expose your ignorance. Such phenomena have been studied from lots of angles. Philosophers of language have developed speech act theory, linguists study the semantics of questions, computer scientists study communication mechanisms, and game theorists have their signaling games. All these perspectives are important – but there is also a foothold for *logic*. This paper will try to show that communication is a typical arena for logical analysis. Logical models help in raising and sometimes solving basic issues not recognized before.

1.2 The puzzle of the Muddy Children

Subtleties of information flow are often high-lighted in puzzles, some with a long history of appeal to broad audiences. A perennial example is Muddy Children:

> *After playing outside, two of three children have mud on their foreheads. They all see the others, but not themselves, so they do not know their own status. Now their Father comes and says: "At least one of you is dirty". He then asks: "Does anyone know if he is dirty?" The children answer truthfully. As this question–answer episode repeats, what will happen?*



Nobody knows in the first round. But upon seeing this, the muddy children will both know in the second round, as each of them can argue as follows.

> "If I were clean, the one dirty child I see would have seen only
> clean children around her, and so she would have known that
> she was dirty at once. But she did not. So I must be dirty, too!"

This is symmetric for both muddy children – so both know in the second round. The third child knows it is clean one round later, after they have announced that. The puzzle is easily generalized to other numbers of clean and dirty children. Many variants are still emerging, as one can check by a simple Internet search.

Puzzles have a serious thrust, as they highlight subtle features of communication beyond simple questions and answers. E.g., consider a putative *Learning Principle* stating that what we hear in public becomes common knowledge. This holds for announcing simple facts – such as the one in Tacitus that, long before international UN peace-keepers, German imperial guards already policed the streets of Rome. But the Principle is not valid in general! In the first round of Muddy Children, the muddy ones both announced the true fact that they did not know their status. But the result of that announcement was not that this ignorance became common knowledge. The announcement rather produced its own falsity, since the muddy children knew their status in the second round. Communicative acts involve *timing* and *information change*, and these may change truth values in complex ways. As we shall see, one of the virtues of logic is that it can help us keep all this straight.

1.3 Logical models of public communication

A logical description of our question-answer episode is easy to give. First, we need to picture the relevant *information states*, after that, we say how they are *updated.*

Answering a question One initial information model for the group *{**Q**, **A**}* of you and me has two states with 'ϕ' ,'$\neg\phi$', with ϕ "this is the road to the Colosseum". We draw these states as points in a diagram. Also, we indicate agents' uncertainties between states. The labeled line shows that ***Q*** cannot distinguish between the two:

$$\phi \;\bullet\!\!-\!\!-\!\!-\!\!-\; Q \;-\!\!-\!\!-\!\!-\!\!\circ\; \neg\phi$$

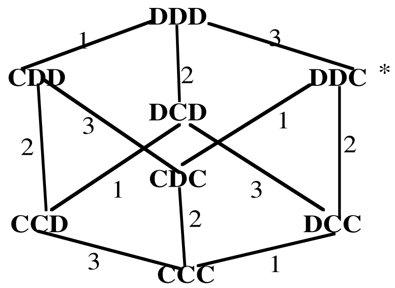
The black dot is an outside marker for the actual world where the agents live. There are no uncertainty lines for ***A***. This reflects the fact that the Roman local ***A*** knows if this is the road to the Colosseum. But ***Q***, though uninformed about the facts, sees that ***A*** knows in each eventuality, and hence he knows that ***A*** knows. This information about other's information is an excellent reason for asking a question.

Next, ***A***'s answer triggers an *update* of this information model. In this simple case, ***A***'s answer eliminates the option *not*-ϕ, thereby changing the initial situation into the following one-point diagram:

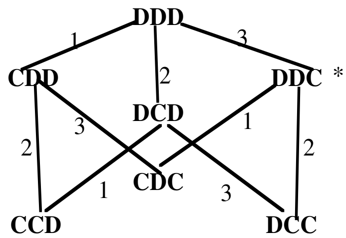
ϕ ●

This picture has only one possible state of the world, where the proposition ϕ holds, and no uncertainty line for anyone. This indicates that ϕ is now common knowledge between you and me. Cognoscenti will recognize where we are heading. Information states are models for the modal logic *S5* in its multi-agent version, and communication consists in actions which change such models. In what follows, we mean by 'knowledge' only: "according to the agent's information".

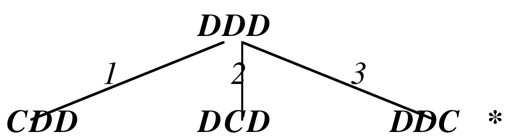
Muddy Children: the movie Here is a video of information updates for Muddy Children. States of the world assign ***D*** (dirty) or ***C*** (clean) to each child: 8 in total. In any of these, a child has one uncertainty. It knows about the others' faces, but cannot distinguish the state from one where its own ***D/C*** value is different.



Updates start with the Father's elimination of the world ***CCC***:



When no one knows his status, the bottom worlds disappear:



The final update is to

DDC *

1.4 General communication

Update by elimination of worlds incompatible with a statement made publicly is a simple mechanism. Human communication in general is very complex, including many other propositional attitudes than knowledge, such as belief or doubt – and logically more challenging phenomena than announcing the truth, such as hiding and cheating. There are two main lines of research here. One is further in-depth analysis of public communication, which turns out to be a quite subtle affair. This will be the main topic of the present paper. The other direction is modeling more complex communicative actions, such as giving answers to questions which others do not hear, or which others overhear, etc. Natural language has a rich vocabulary for all kinds of attitudes toward information, speech acts, secrets, and so on – reflecting our natural proficiency with these. We will discuss more complex models briefly later on. Actually, this might seem a hopeless enterprise, as our behaviour is so diverse and open-ended. But fortunately, there exist simple realistic settings highlighting key aspects, viz. *games* which will also be discussed toward the end.

Some crucial references for this research program are Fagin, Halpern, Moses & Vardi 1995, Gerbrandy 1999, Baltag-Moss & Solecki 1998, and the extensive new mathematical version 2002 of the latter basic reference. Also well-worth reading is van Ditmarsch 2000, which contains a mathematical analysis of all the subtle information passing moves in the well-known parlour game "Cluedo". The present paper builds on these references and others, while also including a number of results by the author over the past few years, mostly unpublished.

2 The basics of update logic

The logic of public information update can be assembled from existing systems. We survey basic epistemic logic and dynamic logic, and then discuss their combination.

2.1 Epistemic logic

Language Epistemic logic has an explicit notation talking about knowledge:

$K_j\ \phi$ agent *j knows that* ϕ

With such a symbolism, we can also analyse further patterns:

$\neg K_j \neg \phi$ (or $<j>\phi$)	agent j considers it *possible that* ϕ
$K_j\, \phi \vee K_j\, \neg\phi$	agent j knows *if* ϕ
$K_j\, \neg\, K_i\, \phi$	j knows that i does not know that ϕ

E.g., in asking a 'normal' question, ***Q*** conveys he does not know if ϕ:

$$\neg K_{\boldsymbol{Q}}\phi \;\;\&\;\; \neg K_{\boldsymbol{Q}}\neg\phi$$

and also that he thinks that ***A*** might know:

$$<\boldsymbol{Q}>\ (K_{\boldsymbol{A}}\phi \vee K_{\boldsymbol{A}}\neg\phi)$$

By answering affirmatively, ***A*** conveys that she knows that ϕ , but she also makes ***Q*** know that ϕ etc., leading to *common knowledge*, which is written as follows:

$$C_{\{Q,\, A\}}\phi$$

Models Models for this epistemic language are of the form

$$\boldsymbol{M} \;=\; (S,\;\; \{\sim_j \mid j \in G\}\ ,\ V)$$

with (a) S a set of worlds, (b) V a valuation function for proposition letters, and (c) for each agent $a \in G$, an *equivalence relation* $\sim_j$ relating worlds s to all worlds that j cannot distinguish from it. These may be viewed as collective information states.

Semantics Next, in these models, an agent *a knows* those propositions that are true in all worlds she cannot distinguish from the current one. That is:

$$\boldsymbol{M},\, s \mid= K_j\, \phi \qquad \text{iff} \qquad \boldsymbol{M},\, t \mid= \phi \quad \textit{for all}\;\; t\;\; \textit{s.t.}\;\; s \sim_j t$$

The related notation $\neg K_j \neg\phi$ or $<j>\phi$ works out to:

$$\boldsymbol{M},\, t \mid= <j>\, \phi \qquad \text{iff} \qquad \boldsymbol{M},\, t \mid= \phi \quad \textit{for some}\;\; t\;\; \textit{s.t.}\;\; s \sim_j t$$

In addition, there are several useful operators of 'group knowledge':

Universal knowledge $E_G\phi$

This is just the conjunction of all formulas $K_j\phi$ for $j\in G$

Common knowledge $C_G\phi$

This says at s that ϕ is true in every state reachable from s through some finite path of uncertainty links for any members of group G

Implicit knowledge $I_G\phi$

This says that ϕ is true in all states which are related to s via the *intersection* of all uncertainty relations $\sim_j$ for $j\in G$

Logic Information models validate an epistemic logic that can describe and automate reasoning with knowledge and ignorance. Here are its major validities:

$K_j(\phi\rightarrow\psi) \rightarrow (K_j\phi\rightarrow K_j\psi)$	*Knowledge Distribution*
$K_j\phi \rightarrow \phi$	*Veridicality*
$K_j\phi \rightarrow K_jK_j\phi$	*Positive Introspection*
$\neg K_j\phi \rightarrow K_j\neg K_j\phi$	*Negative Introspection*

The complete system is multi-*S5*, which serves in two different guises: describing the agents' own explicit reasoning, and describing our reasoning as theorists about them. And here are the required additional axioms for *common knowledge*:

$C_G\,\phi \leftrightarrow \phi\ \&\ E_G\,C_G\,\phi$	*Equilibrium Axiom*
$(\phi\ \&\ C_G\,(\phi\rightarrow E_G\,\phi)) \rightarrow C_G\,\phi$	*Induction Axiom*

The complete logic is also decidable. This is the standard version of epistemic logic.

2.2 Dynamic logic

The usual logic of knowledge by itself can only describe static snapshots of a communication sequence. Now, we must add actions.

Language The language has formulas F and program expressions P on a par:

$F :=$ *propositional atoms* $p, q, r, \ldots \mid \neg F \mid (F\&F) \mid \langle P\rangle F$

$P :=$ *basic actions* $a, b, c, \ldots \mid (P;P) \mid (P\cup P) \mid P^* \mid (F)?$

Semantics This formalism is interpreted over polymodal models

$$\boldsymbol{M} = \langle S,\ \{R_a\}_{a \in A},\ V\rangle$$

which are viewed intuitively as process graphs with states and possible basic transitions. The truth definition explains two notions in one recursion.

$\boldsymbol{M}, s \models \phi$	*ϕ is true at state s*
$\boldsymbol{M}, s_1, s_2 \models \pi$	*the transition from s_1 to s_2 corresponds to a successful execution for the program π*

Here are the inductive clauses:

- $\boldsymbol{M}, s \models p$ iff $s \in V(p)$
  $\boldsymbol{M}, s \models \neg\psi$ iff *not* $\boldsymbol{M}, s \models \psi$
  $\boldsymbol{M}, s \models \phi_1\ \&\ \phi_2$ iff $\boldsymbol{M}, s \models \phi_1$ *and* $\boldsymbol{M}, s \models \phi_2$
  $\boldsymbol{M}, s \models \langle\pi\rangle\phi$ iff *for some* s' *with* $\boldsymbol{M}, s, s' \models \pi$: $\boldsymbol{M}, s' \models \phi$

- $\boldsymbol{M}, s_1, s_2 \models a$ iff $(s_1, s_2) \in R_a$
  $\boldsymbol{M}, s_1, s_2 \models \pi_1 ; \pi_2$ iff *there exists* s_3 *with* $\boldsymbol{M}, s_1, s_3 \models \pi_1$ *and* $\boldsymbol{M}, s_3, s_2 \models \pi_2$
  $\boldsymbol{M}, s_1, s_2 \models \pi_1 \cup \pi_2$ iff $\boldsymbol{M}, s_1, s_2 \models \pi_1$ *or* $\boldsymbol{M}, s_1, s_2 \models \pi_2$
  $\boldsymbol{M}, s_1, s_2 \models \pi^*$ iff *some finite sequence of* π*–transitions in* $\boldsymbol{M}$ *connects* s_1 *with* s_2
  $\boldsymbol{M}, s_1, s_2 \models (\phi)?$ iff $s_1 = s_2$ *and* $\boldsymbol{M}, s_1 \models \phi$

Thus, formulas have the usual Boolean operators, while the existential modality $\langle\pi\rangle\phi$ is a weakest precondition true at only those states where program π can be performed to achieve the truth of ϕ. The program constructions are the usual regular operations of relational composition, Boolean choice, Kleene iteration, and tests for formulas. This system defines standard control operators on programs such as

IF ε THEN π_1 ELSE π_2	$((\varepsilon)? ; \pi_1) \cup ((\neg\varepsilon)? ; \pi_2)$
WHILE ε DO π	$((\varepsilon)? ; \pi)^* ; (\neg\varepsilon)?$

Logic Dynamic logic expresses all of modal logic plus regular relational set algebra. Its complete set of validities is known (cf. Kozen, Harel & Tiuryn 2000):

- All principles of the minimal modal logic for all modalities $[\pi]$
- Computation rules for weakest preconditions:

  $<\pi_1;\pi_2>\phi \leftrightarrow <\pi_1><\pi_2>\phi$

  $<\pi_1\cup\pi_2>\phi \leftrightarrow <\pi_1>\phi \vee <\pi_2>\phi$

  $<\phi?>\psi \leftrightarrow \phi \,\&\, \psi$

  $<\pi^*>\phi \leftrightarrow \phi \vee <\pi><\pi^*>\phi$

- Induction Axiom $\quad (\phi \,\&\, [\pi^*](\phi \rightarrow [\pi]\phi)) \rightarrow [\pi^*]\phi$

The system is also decidable. This property remains also with certain *extensions* of the basic language, such as the program construction $\cap$ of *intersection* – which will return below. Extended modal languages occur quite frequently in applications.

2.3 Dynamic epistemic logic

Analyzing communication requires a logic of knowledge in action, combining epistemic logic and dynamic logic. This may be done in at least two ways.

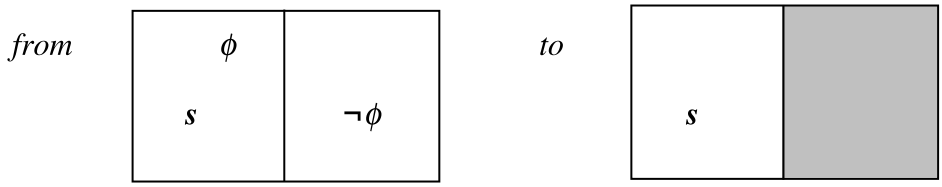
Abstract DEL One can join the languages of epistemic and dynamic logic, and merge the signatures of their models. This yields abstract logics of knowledge and action, cf. Moore 1985 on planning, van Benthem 2001A on imperfect information games. The general logic is the union of epistemic multi-*S5* and dynamic logic. This is a good base for experimenting with further constraints. An example is agents having perfect memory for what went on in the course of communication (cf. Halpern & Vardi 1989). This amounts to an additional commutation axiom

$$K_j[a]\phi \rightarrow [a]K_j\,\phi$$

Abstract *DEL* may be the best setting for general studies of communication.

Concrete update logic In Section 1, public announcement of a proposition ϕ changes the current epistemic model ***M****, s*, with actual world *s,* as follows:

eliminate all worlds which currently do not satisfy ϕ



Thus, we work in a universe whose states are epistemic models – either all of them or just some family – and basic actions are public announcements $A!$ of assertions A from the epistemic language. These actions are *partial functions*. If A is true, then it can be truthfully announced with a unique update. From the standpoint of dynamic logic, this is just one instance of abstract process models, with some epistemic extras. The appropriate logic has combined dynamic-epistemic assertions

$[A!]\,\phi$ "after truthful announcement of A, ϕ holds"

The logic of this system merges epistemic with dynamic logic, with some additions reflecting particulars of our update universe. There is a complete and decidable axiomatization (Plaza 1989, Gerbrandy 1999), with key axioms:

$$\langle A!\rangle p \quad\leftrightarrow\quad A \;\&\; p \quad \textit{for atomic facts } p$$

$$\langle A!\rangle\neg\phi \quad\leftrightarrow\quad A \;\&\; \neg\langle A!\rangle\phi$$

$$\langle A!\rangle\phi\vee\psi \quad\leftrightarrow\quad \langle A!\rangle\phi \vee \langle A!\rangle\psi$$

$$\langle A!\rangle K_i\phi \quad\leftrightarrow\quad A \;\&\; K_i(A \rightarrow \langle A!\rangle\phi)$$

Essentially, these compute preconditions $\langle A!\rangle\phi$ by *relativizing* the postcondition ϕ to A. The axioms can also be stated with the modal box, leading to versions like

$$[A!]K_i\phi \quad\leftrightarrow\quad A \rightarrow K_i(A \rightarrow [A!]\phi)$$

This axiom is like the above law for Perfect Recall. As for common knowledge, the earlier epistemic language needs a little extension, with a *binary* version

$C_G(A,\ \phi)$ common knowledge of ϕ in the submodel defined by A

There is no definition for this in terms of just absolute common knowledge. Having added this feature, we can state the remaining reduction principle

$$\langle A!\rangle\, C_G\phi \quad\leftrightarrow\quad C_G\,(A,\ \langle A!\rangle\phi)$$

These two systems do not exhaust all ways of combining knowledge and action. Van Benthem 1999A sketches a more thoroughly epistemized dynamic logic.

DEL with program constructions Public announcement is just one basic action. Conversation may involve more complex programming of what is said. Saying one thing after another amounts to program composition, choosing one's assertions involves choice, and Muddy Children even involved a guarded iteration:

WHILE 'you don't know your status' DO 'say so'.

The basic logic of public update with the first two constructions is like its version with just basic announcements *A!*, because of the reduction axioms for composition and choice in dynamic logic. But with possible iteration of announcements, the system changes – and even loses its decidability (Baltag, Moss & Solecki 2002).

3 Basic theory of information models

Special model classes Multi-*S5* models can be quite complicated. But there are some subclasses of special interest. For instance, Muddy Children started with a full cube of 3-vectors, with accessibility given as the special equivalence relation

$$X \sim_j Y \qquad \text{iff} \qquad (X)_j - (Y)_j$$

Cube models are studied in algebraic logic (Marx & Venema 1997) for their connections with assignment spaces over first-order models. But the subsequent Muddy Children updates led to submodels of such cubes. These already attain full epistemic generality (van Benthem 1996):

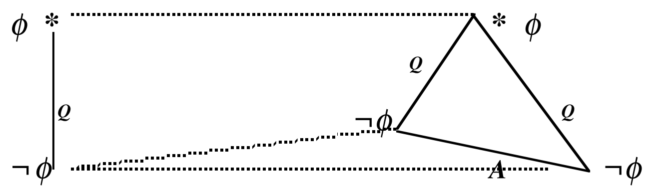
<u>Theorem</u> Every multi-S5 model is representable as a submodel of a cube.

Other special model classes arise in the study of card games (van Ditmarsch 2000).

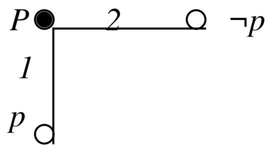
Bisimulation Epistemic and dynamic logic are both standard modal logics (cf. Blackburn, de Rijke, & Venema 2001) with this structural model comparison:

<u>Definition</u> A *bisimulation* between two models $\boldsymbol{M}$, $\boldsymbol{N}$ is a binary relation $\equiv$ between their states m, n such that, whenever $m \equiv n$, then (a) m, n satisfy the same proposition letters, (b1) if $m\ R\ m'$, then there exists a world n' with $n\ R\ n'$ and $m' \equiv n'$, (b2) the same 'zigzag clause' holds in the opposite direction.

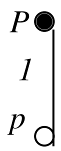
E.g., our question-answer example has a bisimulation with the following variant:



In a natural sense, these are two representations of the same information state. Bisimulation equivalence occurs naturally in update. Suppose that the current model is like this, with the actual world indicated by the black dot:



Note that all three worlds satisfy different epistemic formulas. Now, despite her uncertainty, *1* knows that *p*, and can say this – updating to the model



But this con be contracted via a bisimulation to the one-point model

It is convenient to think of update steps with automatic bisimulation contractions.

Some basic results link bisimulation to truth of modal formulas. For convenience, we restrict attention henceforth to *finite models* – but this can be lifted.

Invariance and definability Consider general models, or those of multi-S5.

<u>Invariance Lemma</u> The following are equivalent:

(a) ***M****, s* and ***N****, t* are connected by a bisimulation

(b) ***M****, s* and ***N****, t* satisfy the same modal formulas

Any model has a bisimilar *unraveled tree model*, but also a smallest *bisimulation contraction* satisfying the same modal formulas. But there is another useful tool:

State Definition Lemma For each model $\boldsymbol{M}$, s there is an epistemic formula β (involving common knowledge) such that the following are equivalent:

(a) $\boldsymbol{N}, t \models \beta$

(b) $\boldsymbol{N}, t$ has a bisimulation $\equiv$ with $\boldsymbol{M}, s$ such that $s \equiv t$

Proof This result is due to Alexandru Baltag (cf. Barwise & Moss 1997). The version and proof given here are from van Benthem 1997, 1998. Consider any finite multi-*S5* model $\boldsymbol{M}, s$. This falls into a number of maximal 'zones' consisting of worlds that satisfy the same epistemic formulas in our language.

Claim 1 There exists a finite set of formulas ϕ_i $(1 \leq i \leq k)$ such that (a) each world satisfies one of them, (b) no world satisfies two of them (i.e., they define a partition of the model), and (c) if two worlds satisfy the same formula ϕ_i, then they agree on all epistemic formulas.

To show this, take any world s, and find 'difference formulas $\delta^{s,t}$ between it and any t which does not satisfy the same epistemic formulas, where s satisfies $\delta^{s,t}$ while t does not. The conjunction of all $\delta^{s,t}$ is a formula ϕ_i true only in s and the worlds sharing its epistemic theory. We may assume the ϕ_i also list all information about the proposition letters true and false throughout their partition zone. We also make a quick observation about uncertainty links between these zones:

\# If any world satisfying ϕ_I is $\sim_a$-linked to a world satisfying ϕ_i, then all worlds satisfying ϕ_I also satisfy $<a>\phi_i$

Next take the following description $\beta_{M,s}$ of $\boldsymbol{M}, s$:

(a) all (negated) proposition letters true at s plus the unique ϕ_i true at $\boldsymbol{M}, s$

(b) common knowledge for the whole group of

- (b1) the disjunction of all ϕ_i
- (b2) all negations of conjunctions $\phi_I \& \phi_j$ $(i \neq j)$
- (b3) all implications $\phi_i \rightarrow <a>\phi_j$ for which situation # occurs
- (b4) all implications $\phi_i \rightarrow [a]\bigvee\phi_j$ where the disjunction runs over all situations listed in the previous clause.

<u>Claim 2</u> $\quad \boldsymbol{M}, s \mid= \beta_{M, s}$

<u>Claim 3</u> $\quad$ If $\boldsymbol{N}, t \mid= \beta_{M, s}$, then there is a bisimulation between $\boldsymbol{N}, t$ and $\boldsymbol{M}, s$

To prove Claim 3, let $\boldsymbol{N}, t$ be any model for $\beta_{M, s}$. The ϕ_i partition $\boldsymbol{N}$ into disjoint zones Z_i of worlds satisfying these formulas. Now relate all worlds in such a zone to all worlds that satisfy ϕ_i in the model $\boldsymbol{M}$. In particular, t gets connected to s. We must check that this connection is a bisimulation. The atomic clause is clear from an earlier remark. But also, the zigzag clauses follow from the given description. (a) Any $\sim_a$-successor step in $\boldsymbol{M}$ has been encoded in a formula $\phi_i \rightarrow \langle a \rangle \phi_j$ which holds everywhere in $\boldsymbol{N}$, producing the required successor there. (b) Conversely, if there is no $\sim_a$-successor in $\boldsymbol{M}$, this shows up in the limitative formula $\phi_i \rightarrow [a] \vee \phi$, which also holds in $\boldsymbol{N}$, so that there is no 'excess' successor there either. ■

The Invariance Lemma says bisimulation has the right fit with the modal language. The State Definition Lemma says each semantic state can be characterized by one epistemic formula. E.g., consider the two-world model for our question-answer episode. Here is an epistemic formula which defines its ϕ-state up to bisimulation:

$$\phi \;\&\; C_{\{Q, A\}} ((K_A\phi \vee K_A\neg\phi) \;\&\; \neg K_Q\phi \;\&\; \neg K_Q\neg\phi)$$

This allows us to switch, in principle, between semantic accounts of information states as models $\boldsymbol{M}, s$ and syntactic ones in terms of complete defining formulas. There is more to this than just technicality. For instance, syntactic approaches have been dominant in related areas like belief revision theory, where information states are not models but syntactic theories. It would be good to systematically relate syntactic and semantic approaches to update, but we shall stay semantic here.

Respectful and safe operations The above also constrains epistemic update operations O. These should *respect bisimulation*:

If $\boldsymbol{M}, s$ and $\boldsymbol{N}, t$ are bisimilar, so are their values $O(\boldsymbol{M}, s)$ and $O(\boldsymbol{N}, t)$

<u>Fact</u> Public update respects bisimulation.

<u>Proof</u> Let $\equiv$ be a bisimulation between $\boldsymbol{M}, s$ and $\boldsymbol{N}, t$. Consider their submodels $\boldsymbol{M}|\phi, s$, $\boldsymbol{N}|\phi, t$ after public update with ϕ. The *restriction* of $\equiv$ to these is still a

bisimulation. Here is the zigzag clause. Suppose some world w has an $\sim_i$-successor v in $\boldsymbol{M}|\phi, s$. This same v is still available in the other model: it remained in $\boldsymbol{M}$ since it satisfied ϕ. But then v also satisfied ϕ in $\boldsymbol{N}, t$, because of the Invariance Lemma for the bisimulation $\equiv$ – and so it stayed in the updated model $\boldsymbol{N}|\phi, t$, too. ■

Many other proposed update operations respect bisimulations (cf. also Hollenberg 1998 on process algebra). Finally, bisimulation also works for dynamic logic – but with a new twist (van Benthem 1996). Intertwined with invariance for formulas ϕ, one must show that the zigzag clauses go through for all regular program constructions: not just the atomic R_a, but each transition relation $[[\pi]]$:

<u>Fact</u> Let $\equiv$ be a bisimulation between two models $\boldsymbol{M}, \boldsymbol{M}'$, with $s \equiv s'$.

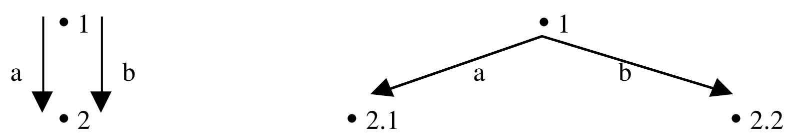
(i) s, s' verify the same formulas of propositional dynamic logic

(ii) if $s\ [[\pi]]^{\boldsymbol{M}}\ t$, then there exists t' with $s'\ [[\pi]]^{\boldsymbol{M}'}\ t'$ and $s' \equiv t'$

This observation motivates this notion of invariance for program operations

<u>Definition</u> An operation $O(R_1, ..., R_n)$ on programs is *safe for bisimulation* if, whenever $\equiv$ is a relation of bisimulation between two models for their transition relations $R_1, ..., R_n$, then it is also a bisimulation for $O(R_1, ..., R_n)$.

The core of the above program induction is that the three regular operations $;\ \cup\ ^*$ of PDL are safe for bisimulation. By contrast, program *intersection* is not safe:



There is an obvious bisimulation with respect to a, b – but zigzag fails for $R_a \cap R_b$.

After indulging in this technical extravaganza, it is time to return to communication. In fact, the Muddy Children puzzle highlights a whole agenda of further questions. We already noted how its specific model sequence is characteristic for the field. But in addition, it raises many further central issues, such as

(a) the benefits of internal group communication

(b) the role of iterated assertion

(c) the interplay of update and inference in reasoning.

We will look into these as we go. But we start with an issue which we already noted: the putative 'learning principle' that was refuted by Muddy Children.

4 What do we learn from a statement?

Specifying speech acts Update logic may be considered a sequel to dynamic speech act theories, which originated in philosophy, and then partly migrated to computer science (cf. Wooldridge 2002). Earlier accounts of speech acts often consist in formal specifications of preconditions and postconditions of successful assertions, questions, or commands. Some of these insights are quite valuable, such as those concerning assertoric force of assertions. E.g., in what follows, we will assume, in line with that tradition, that normal cooperative speakers may only utter statements which they know to be true. Even so, what guarantees that the specifications are correct? E.g., it has been said that answers to questions typically produce common knowledge of the answer. But Muddy Children provided a counter-example to this putative 'Learning Principle'. Logical tools help us get clearer on pitfalls and solutions. The learning problem is a good example.

Persistence under update Public announcement of atomic facts p makes them common knowledge, and the same holds for other types of assertion. But, as we noted in Section 1, not all updates with ϕ result in common knowledge of ϕ! A simple counter-example is this. In our question-answer case, let ***A*** say truly

$p \ \& \ \neg K_Q p$ "*p*, but you don't know it"

This very utterance removes ***Q***'s lack of knowledge about the fact p, and thereby makes its own assertion false! Ordinary terminology is misleading here:

> *learning that* ϕ is ambiguous between: ϕ *was* the case, before the announcement, and ϕ *is* the case – after the announcement.

The explanation is that statements may change truth value with update. For worlds surviving in the smaller model, factual properties do not change, but epistemic properties may. This raises a general logical issue of *persistence under update*:

> Which forms of epistemic assertion remain true at a world whenever other worlds are eliminated from the model?

These are epistemic assertions which, when publicly announced to a group, will always result in common knowledge. Examples are atomic facts p, and knowledge-free assertions generally, knowledge assertions K_ip, ignorance assertions $\neg K_ip$.

New kinds of preservation results Here is a relevant result from modal logic (cf. Andréka, van Benthem & Németi 1998):

Theorem The epistemic formulas *without common knowledge* that are preserved under submodels are precisely those definable using literals p, $\neg p$, conjunction, disjunction, and K_i-operators.

Compare universal formulas in first-order logic, which are just those preserved under submodels. The obvious conjecture for the epistemic language with common knowledge would allow arbitrary C–operators as well. But this result is still open, as lifting first-order model theory to modal fixed-point languages seems non-trivial.

Open Question Which formulas of the full epistemic language with common knowledge are preserved under submodels?

In any case, what we need is not really full preservation under submodels, but rather preservation under 'self-defined submodels':

When we *restrict* a model to those of its worlds which satisfy ϕ,
then ϕ should hold throughout the remaining model,
or in terms of an elegant validity: $\phi \rightarrow (\phi)^{\phi}$

Open Question Which epistemic formulas imply their self-relativization?

For that matter, which first-order formulas are preserved in this self-fulfilling sense? Model-theoretic preservation questions of this special form seem new.

A non-issue? Many people find this particular issue annoying. Non-persistence seems a side-effect of infelicitous wording. E.g., when ***A*** said "*p*, but you don't know it", she should just have said "*p*", keeping her mouth shut about my mental state. Now, the Muddy Children example is not as blatant as this. And in any case, dangers in timing aspects of what was true before and is true after an update are no more exotic than the acknowledged danger in computer science of confusing states of a process. Dynamic logics were developed precisely to keep track of that.

Let's stop fencing: *can* we reword any message to make the non-persistence go away? An epistemic assertion A defines a set of worlds in the current model $\boldsymbol{M}$. Can we always find an equivalent persistent definition? This would be easy if each world has a simple unique factual description, like hands in card games. But even without assuming thisthere is a method that works, at least locally:

Fact In each model, every public announcement has a persistent equivalent.

Proof Without loss of generality, assume we are working only with bisimulation-contracted models which are also totally connected: no isolated components. Let w be the current world in model $\boldsymbol{M}$. Let j publicly announce A, updating to the submodel $\boldsymbol{M}|A$ with domain $A^* = \{s \in \boldsymbol{M} \mid \boldsymbol{M}, s \models A\}$. If this is still $\boldsymbol{M}$ itself, then the announcement "True" is adequate, and persistent. Now suppose A^* is not the whole domain. Our persistent assertion consist of two disjuncts:

$$\Delta \vee \Sigma$$

First we make Δ. Using the proof of the State Definition Lemma of Section 3, this is an epistemic definition for A^* in $\boldsymbol{M}$ formed by describing each world in it up to bisimulation, and then taking the disjunction of these.

Now for Σ. Again using the mentioned proof, write a formula which describes $\boldsymbol{M}|A$ up to bisimulation. For concreteness, this had a common knowledge operator over a plain epistemic formula describing the pattern of states and links, true everywhere in the model $\boldsymbol{M}|A$. No specific world description is appended, however.

Next, $\Delta \vee \Sigma$ is common knowledge in $\boldsymbol{M}|A$, because Σ is. But it also picks out the right worlds in $\boldsymbol{M}$. Clearly, any world in A^* satisfies its own disjunct of Δ. Conversely, suppose any world t in $\boldsymbol{M}$ satisfies $\Delta \vee \Sigma$. If it satisfies some disjunct of Δ, then it must then be in A^* by the bisimulation-minimality of the model. Otherwise, $\boldsymbol{M}, t$ satisfies Σ. But then by connectedness, every world in $\boldsymbol{M}$ satisfies Σ, and in particular, given the construction of Σ, there must be a bisimulation between $\boldsymbol{M}$ and $\boldsymbol{M}|A$. But this contradicts the fact that the update was genuine. ■

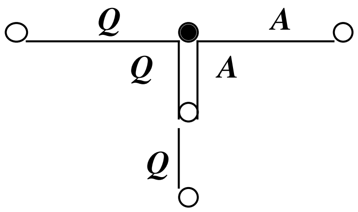
Of course, this recipe for phrasing your assertions is ugly, and not recommended! Moreover, it is local to one model, and does not work uniformly. Recall that, depending on group size, muddy children may have to repeat the same ignorance statement any number of times before knowledge dawns. If there were one uniform persistent equivalent for that statement, the latter's announcement would lead to common knowledge after some fixed finite stage.

5 Internal communication in groups

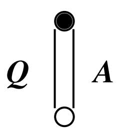
The best we can The muddy children might just tell each other what they see, and common knowledge of their situation is reached at once. The same holds for card players telling each other their hands. Of course, life is civilized precisely because we do not 'tell it like it is'. Even so, there is an issue of principle what agents in a group can achieve by maximal communication. Consider two epistemic agents that find themselves in some collective information state $\boldsymbol{M}$, at some actual situation $\boldsymbol{s}$. They can tell each other things they know, thereby cutting down the model to smaller sizes. Suppose they wish to be maximally cooperative:

> What is the best correct information they can give via successive updates
> – and what does the resulting collective information state look like?

E.g., what is the best that can be achieved in the following model?



Geometrical intuition suggests that this must be:



This is correct! First, any sequence of mutual updates in a finite model must terminate in some minimal domain which can no longer be reduced. This is reached when everything each agent knows is already common knowledge: i.e., it holds in every world. But what is more, this minimal model is *unique*, and we may call it the ‘communicative core’ of the initial model. Here is an explicit description, proved in van Benthem 2000:

<u>Theorem</u> Each model has a communicative core, viz. the set of worlds that are reachable from the actual world via all uncertainty links.

<u>Proof</u> For convenience, consider a model with two agents only. The case with more than two agents is an easy generalization of the same technique.

First, agents can reach this special set of worlds as follows. Without loss of generality, let all states t in the model satisfy a unique defining formula δ_t as in Section 3 – or obtained by an ad-hoc argument. Agent *1* now communicates all he knows by stating the *disjunction* $\vee\delta_t$ for all worlds t he considers indistinguishable from the actual one. This initial move cuts the model down to the actual world plus all its $\sim_1$-alternatives. Now there is a small technicality. The resulting model need no longer satisfy the above unique definability property. The update may have removed worlds that distinguished between otherwise similar options. But this is easy to remedy by taking the *bisimulation contraction*. Next, let *2* make a similar strongest assertion available to her. This cuts the model down to those worlds that are also $\sim_2$-accessible from the actual one. After that, everything any agent knows is common knowledge, so further statements have no informative effect.

Next, suppose agents reach a state where further announcements have no effect. Then the following implications hold for all ϕ: $K_1\phi \rightarrow C_{\{1,\ 2\}}\phi$, $K_2\phi \rightarrow C_{\{1,\ 2\}}\phi$. Again using defining formulas, this means *1, 2* have the same alternative worlds. So, these form a *subset* of the above core. But in fact, all of it is preserved. An agent can only make statements that hold in all of its worlds, as it is included in his information set, Therefore, the whole core survives each episode of public update, and by induction, it survives all of them. ■

A corollary of the preceding proof is this:

<u>Fact</u> Agents need only 2 rounds of communication to get to the core.

In particular, there is no need for repetitions by agents. E.g., let *1* truly say A (something he knows in the actual world), note the induced public update, and then say B (which he knows in the new state). Then he might just as well have asserted A & $(B)^A$ straightaway: where $(B)^A$ is the *relativization* of B to A (cf. Section 6).

Incidentally, a two-step solution to the initial example of this section is the following rather existentialist conversation:

Q sighs: "I don't know"
A sighs: "I don't know either"

It does not matter if you forget details, because it also works in the opposite order.

The communicative core is the actual world plus every world connected to it by the intersection of all uncertainty relations. This is the range used in defining *implicit knowledge* for a group of agents in Section 2.1. Thus, maximal communication turns implicit knowledge of a group into common knowledge. As a slogan, this makes sense, but there are subtleties. It may be implicit knowledge that none of us know where the treasure is. But once the communicative core is all that is left, the location of the treasure may be common knowledge. Compare the difference between quantifier restriction and relativization. Implicit knowledge $I_G\phi$ looks only at worlds in the communication core CC, but it then evaluates the formula ϕ from each world there in the whole model. By contrast, internal evaluation in just the core is like evaluating totally relativized statements $(\phi)^{CC}$ in the model.

Another technicality is that the relevant *intersection* of relations, though keeping the logic decidable, is no longer safe for bisimulation in the sense of Section 4. Adding it to the language leads to a genuinely richer epistemic logic, for which some of the earlier model theory would have to be redone.

Planning assertions This section shows a shift in interest. Update logics can be used to analyze given assertions, but they can also be used to plan assertions meeting certain specifications. A more complex example is the following puzzle from a mathematical Olympiad in Moscow (cf. van Ditmarsch 2002):

> *7* cards are distributed among *A, B, C*. *A* gets *3*, *B* gets *3*, C gets *1*.
> How should *A, B* communicate publicly, in hearing of *3* , so that
> they find out the precise distribution of the cards while *C* does not?

There are solutions here – but their existence depends on the number of cards. This question may be seen as a generalization of the preceding one. How can a subgroup of all agents communicate maximally, while keeping the rest of the group as much in the dark as possible? Normally, this calls for hiding, but it is interesting to see – at least to politicians, or illicit lovers – that some of this can be achieved publicly. This sort of planning problem is only beginning to be studied.

Here we just observe an analogy with computer science. The dynamic epistemic logic of Section 2.3 is like program logics manipulating correctness assertions

$\phi \rightarrow [A!]\ \psi$ if precondition ϕ holds, then saying A always leads to a state where postcondition ψ holds.

Such triples may be looked at in different ways. Given an assertion, one can analyze its preconditions and postconditions, as we did for questions and answers. This is program analysis. Or, with precondition ϕ and assertion $A!$, we can look for their strongest postcondition ψ. perhaps, common knowledge of A. But there is also program synthesis. Given a precondition ϕ and postcondition ψ, we can look for an assertion $A!$ guaranteeing the transition. Conversation planning is like this.

6 Public update as relativization

This technical intermezzo (van Benthem 1999B) joins forces with standard logic.

Semantic and syntactic relativization Here is a simple fact. Announcing A amounts to a logical operation of *semantic relativization*

from a model $\boldsymbol{M}, s$ to the definable submodel $\boldsymbol{M}|A, s$.

This explains all behaviour so far – while raising new questions. For a start, in the new model, we can again evaluate formulas that express knowledge and ignorance of agents, in the standard format $\boldsymbol{M}|A, s \models \phi$. In standard logic, this may also be described via *syntactic relativization* of the formula ϕ by the update assertion A:

<u>Relativization Lemma</u> $\boldsymbol{M}|A, s \models \phi$ iff $\boldsymbol{M}, s \models (\phi)^A$

This says we can either evaluate our assertions in a relativized model, or equivalently, their relativized versions in the original model. For convenience, we will assume henceforth that relativization is defined so that $(\phi)^A$ always implies ϕ. For the basic epistemic language, this goes via the following recursion:

$$(p)^A = A \,\&\, p$$

$$(\neg\phi)^A = A \,\&\, \neg\, (\phi)^A$$

$$(\phi\vee\psi)^A = (\phi)^A \vee (\psi)^A$$

$$(K_i\phi)^A = A \,\&\, K_i(A \rightarrow (\phi)^A)$$

In this definition, one immediately recognizes the above axioms for public update. Whether this works entirely within the language of epistemic announcements depends on its strength. E.g., relativization was less straightforward with common knowledge, as no syntactic prefix $'A \rightarrow \ldots'$ or $'A \& \ldots'$ on absolute operators C_G does the job. But one can extend epistemic logic with a binary restricted common knowledge operator. Actually, dynamic logic is better behaved in this respect.

<u>Fact</u> Dynamic logic is closed under relativization.

<u>Proof</u> In line with the usual syntax of the system, we need a double recursion over formulas and programs. For formulas, the clauses are all as above, while we add

$$([\pi]\phi)^A \quad = \quad [(\pi)^A](\phi)^A$$

For programs, here are the recursive clauses that do the job:

$$
\begin{aligned}
(R\ ;\ S)^A &\quad = \quad (R)^A\ ;\ (S)^A \\
(R \cup S)^A &\quad = \quad (R)^A \cup (S)^A \\
((\phi)?)^A &\quad = \quad (A)?\ ;\ (\phi)?\,, \\
(\pi^*)^A &\quad = \quad ((A)?\ ;\ (\pi)^A)^* \qquad \blacksquare
\end{aligned}
$$

Clearly, common knowledge $C_G\phi$ may be viewed as a dynamic logic formula

$$[(\cup\{i \mid i \in G\})]\ \phi$$

Therefore, we can get a natural relativization for epistemic logic by the above Fact, by borrowing a little syntax from dynamic logic.

General logic of relativization Stripped of its motivation, update logic is an axiomatization of one model-theoretic operation, viz. relativization. There is nothing specifically modal about this. One could ask for a complete logic of relativizations $(\phi)^A$ in first-order logic, as done for *substitutions* $[t/x]\phi$ in Marx & Venema 1997.

<u>Open Question</u> What is the complete logic of relativization in first-order logic?

At least we may observe that there is more than the axioms listed in Section 3.3. For instance, the following additional fact is easy to prove:

Associativity $((A)^B)^C$ is logically equivalent to ${}_A((B)^C)$

In our update logic, performing two relativizations corresponds to performing two consecutive updates. Thus Associativity amounts to the validity of

$$[A!\,;\,B!]\,\phi \leftrightarrow [([A!]B)!]\,\phi$$

Why was this not on the earlier list of the complete axiom system? The answer is a subtlety. That axiom system does indeed derive every valid formula. But it does so without being *substitution-closed*. In particular, the above basic axiom for atoms

$$\langle A!\rangle p \leftrightarrow A \,\&\, p$$

fails for arbitrary formulas ϕ. Define the *substitution core* of update logic as those schemata all of whose substitution instances are valid formulas. Associativity belongs to it, but it is not derivable schematically from the earlier axiom system.

Open Question Axiomatize the substitution core of public update logic.

There are also interestingly invalid principles, witness the discussion of persistence in Section 4. Announcing a true statement "*p*, but you don't know it" invalidates itself. More technically, even when $p \,\&\, \langle 1\rangle\neg p$ holds, its self-relativization

$$(p \,\&\, \langle 1\rangle\neg p)^{p \,\&\, \langle 1\rangle\neg p} \quad = \quad p \,\&\, \langle 1\rangle\neg p \;\;\&\; \langle 1\rangle(\,p \,\&\, \langle 1\rangle\neg p \,\&p)$$

is a contradiction. Thus some assertions are self-refuting when announced, and the following pleasing principle is not part of a general logic of relativization:

$\phi \rightarrow (\phi)^\phi$ holds only for special assertions ϕ

We will look at some further issues in the logic of relativization in Section 7, including iterated announcement and its connections with fixed-point operators.

Excursion: richer systems of update In standard logic, relativization often occurs together with other operations, such as *translation of predicates* – e.g., in the notion of *relative interpretation* of one theory into another. Likewise, the above connection extends to more sophisticated forms of epistemic update (cf. Section 9). For instance, when a group hears that a question is asked and answered, but only a subgroup gets the precise answer, we must use a new operation of *arrow elimination,* rather than world elimination. More precisely,

all arrows are removed for all members of that subgroup between
those zones of the model that reflect different exhaustive answers.

Arrow elimination involves substitution of new accessibility relations for the current ones. E.g., when the question "ϕ?" is asked and answered, the uncertainty relations $\sim_i$ for agents i in the informed subgroup are replaced by the union of relations

$$(\phi)? \,;\, \sim_I ; (\phi)? \cup (\neg\phi)? \,;\, \sim_i ; (\neg\phi)?$$

But this is just translation of the old binary relation $\sim_i$ into a new definable one.

Next on this road, there are more complex 'product updates' – which correspond to those interpretations between theories which involve construction of new definable objects, like when we embed the rationals into the integers using ordered pairs. Axioms for update logics will then still axiomatize parts of the meta-theory of such general logical operations. Thus, progressively more complex notions of update correspond to more sophisticated theory relations from standard logic.

Finally, relativization suggests a slightly different view of eliminative update. So far, we discarded old information states. But now, we can keep the old information state, and perform 'virtual update' via relativized assertions. Thus, the initial state already contains all possible future communicative developments. Another take on this at least keeps the old models around, doing updates with *memory*. There are also independent reasons for maintaining some past history in our logic, having to do with public updates which refer explicitly to the 'epistemic past', such as:

"what you said, I knew already".

See van Benthem 2002A, and also Section 9 below, for more concrete examples.

7 Repeated announcement and limit behaviour

'Keep talking' In the Muddy Children scenario, an assertion of ignorance was repeated until it could no longer be made truly. In the given model, the statement was *self-defeating*: when repeated iteratively, it reaches a stage where it is not true anywhere. Of course, self-defeating ignorance statements lead to something good for us, viz. knowledge. There is also a counterpart to this limit behaviour: iterated announcement of *self-fulfilling* statements makes them common knowledge. This happened in one step with factual assertions and others in Section 4. More subtle cases are discussed in van Benthem 2002B, viz. repeated joint assertions of rationality by players in a strategic game, saying that one will only choose actions that may be best possible responses to what the others do. These may decrease the set of available strategy profiles until a 'best zone' is reached consisting of either a Nash equilibrium, or at least some rationalizable profiles to be in.

Limits and fixed-points Repeated announcement of rationality by two players *1, 2* has the following form, which we take for granted here without motivation:

$$JR: \quad \langle 1 \rangle B_1 \wedge \langle 2 \rangle B_2$$

Where the proposition letter B_1 says that *i*'s action in the current world is a *best response for i* to what the opponent is playing here. It can be shown that any finite game matrix has entries (worlds in the corresponding epistemic model) in a loop

$$x_1 \models B_1 \sim_2 x_2 \models B_2 \sim_1 x_3 \models B_1 \sim_2 \ldots \sim_1 x_1 \models B_1$$

Repeated announcement of joint rationality *JR* may keep removing worlds, as each announcement may remove worlds satisfying a B_i on which one conjunct depended. But clearly, whole loops of the kind described remain all the time, as they form a kind of mutual protection society. Thus, we have a first

<u>Fact</u> Strong Rationality is self-fulfilling on finite game matrix models.

The technical connection with fixed-points suggests extending basic update logic with *fixed-point operators*. This is like extending modal or dynamic logic to the so-called *μ–calculus*, whose syntax provides smallest fixed-point definitions of the form $\mu p \bullet \phi(p)$ and greatest ones of the form $\nu p \bullet \phi(p)$. Stirling 1999 has details on the μ–calculus, Ebbinghaus & Flum 1995 on more general fixed-point logics.

We explore this a bit, as there are some tricky but nice issues involved (for details, cf. the reference). For a start, we can prove this

<u>Fact</u> The stable set of worlds reached via repeated announcement of *JR* is defined inside the original full game model by the greatest fixed-point formula $\nu p \bullet (\langle \mathbf{E} \rangle (B_E \wedge p) \wedge \langle \mathbf{A} \rangle (B_A \wedge p))$

Iterated announcement in dynamic logic In any model $\boldsymbol{M}$, we can keep announcing any formula ϕ, until we reach a fixed-point, perhaps the empty set:

$$\#(\phi, \boldsymbol{M})$$

E.g., *self-fulfilling* formulas ϕ in $\boldsymbol{M}$ become common knowledge in $\#(\phi, \boldsymbol{M})$:

$$\phi \to (C_G\phi)^{\#(\phi, \boldsymbol{M})}$$

What kind of fixed-point are we computing here? Technically $\#(\phi, \boldsymbol{M})$ arises by continued application of this function, taking intersections at limit ordinals:

$$F_{M,\phi}(X) = \{s \in X \mid \boldsymbol{M}|X, s \models \phi\}$$

with $\boldsymbol{M}|X$ the restriction of $\boldsymbol{M}$ to the set X

The map F is *not monotone*, and the usual theory of fixed-points does not apply. The reason is the earlier fact that statements ϕ may change truth value when passing from $\boldsymbol{M}$ to submodels $\boldsymbol{M} \mid X$. In particular, we do not recompute stages inside one unchanging model, as in the normal semantics of greatest fixed-point formulas $\nu p\bullet\ \phi(p)$, but in ever smaller models, changing the range of the modal operators. Thus we mix fixed-point computation with *relativization* (cf. Section 6). Despite F's non-monotonicity, iterated announcement is a fixed-point procedure of sorts:

<u>Fact</u> The iterated announcement limit is an *inflationary* fixed point.

<u>Proof</u> Take any ϕ, and relativize it to a fresh proposition letter p, yielding

$$(\phi)^p$$

Here p need not occur positively (it becomes negative when relativizing positive K–operators). Now the obvious epistemic Relativization Lemma says that

$$\boldsymbol{M}, s \models (\phi)^p \quad \text{iff} \quad \boldsymbol{M}/X, s \models \phi$$

Therefore, the above definition of $F_{M,\phi}(X)$ as $\{s \in X \mid \boldsymbol{M}|X, s \models \phi\}$ equals

$$\{s \in \boldsymbol{M} \mid \boldsymbol{M}, s \models (\phi)^p\} \cap X$$

This computes a greatest *inflationary fixed-point* (Ebbinghaus & Flum 1995). ■

But then, why did iterated announcement of *JR* produce an ordinary greatest fixed-point? The above update map $F_{M,\phi}(X)$ *is* monotone with special sorts of formulas:

<u>Fact</u> $F_{M,\phi}(X)$ is monotone for *existential modal formulas*.

The reason is that such formulas are preserved under model extensions, making their *F* monotone for set inclusion: cf. the related preservation issues in Section 4.

Excursion: comparing update sequences Update logic is subtle, even here. What happens when we compare different repeated announcements of rationality that players could make? Van Benthem 2002B considers a weaker assertion *WR* which follows from *JR*. Does this guarantee that their limits are included:

$$\#(SR, \boldsymbol{M}) \subseteq \#(WR, \boldsymbol{M})?$$

The general answer is negative. Making weaker assertions repeatedly may lead to incomparable results. An example are this formula ϕ and its consequence ψ:

$$\phi = p \wedge (\langle\rangle\neg p \rightarrow \langle\rangle q)$$

$$\psi = \phi \vee (\neg p \wedge \neg q)$$

In the following epistemic model, the update sequence for ϕ stops in one step with the world ***1***, whereas that for ψ runs as follows: ***{1, 2, 3}***, ***{1, 2}***, ***{2}***.

$$\begin{array}{ccccc} \boldsymbol{1} & \text{\_\_\_\_\_\_\_\_\_\_} & \boldsymbol{2} & \text{\_\_\_\_\_\_\_\_\_\_} & \boldsymbol{3} \\ p, \neg q & & \neg p, \neg q & & \neg p, q \end{array}$$

But sometimes, things work out.

<u>Fact</u> If an *existential* ϕ implies ψ in $\boldsymbol{M}$, then $\#(\phi, \boldsymbol{M}) \subseteq \#(\psi, \boldsymbol{M})$.

<u>Proof</u> We always have the inclusion

$$T^{\alpha}{}_{\phi}(\boldsymbol{M}) \subseteq T^{\alpha}{}_{\psi}(\boldsymbol{M})$$

The reason for this is the following implication:

if $X \subseteq Y$, then $F_{M,\phi}(X) \subseteq F_{M,\psi}(Y)$

For, if $\boldsymbol{M}|X,\ s\ |= \phi$ and $s \in X$, then $s \in Y$ and also $\boldsymbol{M}|Y,\ s\ |= \phi$ – by the modal *existential* form of ϕ. But then $\boldsymbol{M}|Y,\ s\ |= \psi$, by our valid implication. ■

One more type of fixed point! Iterated announcement can be described by the finite iteration * of dynamic logic (cf. Section 2.2). This extension is studied in Baltag–Solecki–Moss 2002, which shows that formulas of the form

$$<(A!)^*> C_G A$$

are not definable in the modal μ–calculus. Still, it is well-known that formulas

$$[(A!)^*]\phi$$

with program iteration of this sort are definable with greatest fixed-point operators

$$\nu p \bullet \phi \wedge [A!]p$$

But these cannot be analyzed in the earlier style, as they involve relativizing p to A, rather than the more tractable A to p, as in our analysis of repeated announcement.

8 Inference versus update

Dynamic inference Standard epistemic logic describes inference in unchanging information models. But the current literature also has a more lively notion following the dynamics of update (cf. van Benthem 1996):

> Conclusion ϕ *follows dynamically* from premises $P_1, ..., P_k$ if
> after updating any information state with public announcements
> of the successive premises, all worlds in the end state satisfy ϕ.

In terms of dynamic-epistemic logic, the following implication must be valid:

$$[P_1! ; \ldots ; P_k!]\ C_G\phi$$

This notion behaves differently from standard logic in its premise management:

> *Order of presentation matters*
> Conclusions from A, B need not be the same as from B, A:
> witness $\neg Kp$, p (consistent) versus $p, \neg Kp$ (inconsistent)

Multiplicity of occurrence matters
$\neg Kp\&p$ has different update effects from $(\neg Kp \,\&\, p) \,\&\, (\neg Kp \,\&\, p)$

Adding premises can disturb conclusions
$\neg Kp$ implies $\neg Kp$ – but $\neg Kp, p$ does not imply $\neg Kp$.

By contrast, the structural rules of classical logic say precisely that order, multiplicity, and overkill does not matter. Nevertheless, there is a description.

Structural rules and representation Van Benthem 2001C provides three modified structural rules that are valid for dynamic inference as defined above:

Left Monotonicity $X \Rightarrow A$ implies $B, X \Rightarrow A$
Cautious Monotonicity $X \Rightarrow A$ and $X, Y \Rightarrow B$ imply $X, A, Y \Rightarrow B$
Left Cut $X \Rightarrow A$ and $X, A, Y \Rightarrow B$ imply $X, Y \Rightarrow B$

Moreover, the following completeness result holds:

Theorem The structural properties of dynamic inference are characterized completely by Left Monotonicity, Cautious Monotonicity, and Left Cut.

The core of the proof is a representation argument showing that any abstract finite tree model for modal logic can be represented up to bisimulation in the form:

Worlds w go to a family of epistemic models $\boldsymbol{M}_w$
Basic actions a go to suitable epistemic announcements $(\phi_a)!$

This suggests that public update is a quite general process, which can encode arbitrary processes in the form of 'conversation games'.

Inference versus update Here is amore general moral. Logic has two different inferential processes. The first is *ordinary inference*, related to implications $A \rightarrow B$. This stays inside one fixed model. The second process is a model-jumping *inference under relativization*, related to the earlier formulas $[A!]B$. Both seem interesting. Even so, one empirical issue remains. The muddy children *deduced* a solution: they did not draw update diagrams. What is going on inside our heads?

9 The wider world

As stated in Section 1, update analysis has two main directions: increased coverage by means of new models and mechanisms, and increased in-depth understanding of the logics that we already have. This paper has concentrated on the latter, hoping to show the interest of logical issues in communication. In this Section, the reader gets a lightning tour of what she has missed.

Keeping track of past assertions Some puzzles involve reference to past states, with people saying things like "What you said did not surprise me" (McCarthy 2002). This says that they knew at the previous state, calling for a further update there. To accomplish this, we need to maintain a stack of past updates, instead of just performing them and trashing previous stages. In the limit, as also mentioned earlier, this richer framework might also include protocol information about the sort of communication that we are in.

Privacy and hiding The next stage of complexity beyond public communication involves hiding information, either willfully, or through partial observation. Here is about the simplest example, first stated in the glossy brochure Spinoza 1998:

> *We have both just drawn a closed envelope. It is common knowledge between us that one envelope holds an invitation to a lecture on logic, the other to a wild night out in Amsterdam. We are both ignorant of the fate in store for us! Now I open my envelope, and read the contents, without showing them to you. Yours remains closed. Which information has passed exactly because of my action? I certainly know now which fate is in store for me. But you have also learnt something, viz.* that I know – *though not what I know. Likewise, I did not just learn what is in my envelope. I also learnt something about you, viz. that you know that I know. The latter fact has even become common knowledge between us. And so on. What is a general principle behind this?*

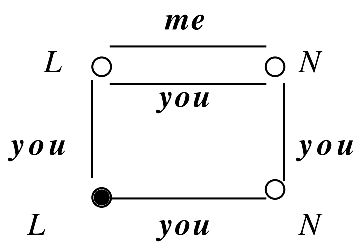
The initial information state is a familiar one of collective ignorance:

$$L \;\bullet\; \overset{\boldsymbol{me}}{\underset{\boldsymbol{you}}{=\!=\!=\!=\!=}} \;\circ\; N$$

The intuitive update just removes my uncertainty link – while both worlds remain available, as they are needed to model your continued ignorance of the base fact:

$$L \;\bullet\; \overset{\boldsymbol{you}}{\text{———}} \;\circ\; N$$

Such updates occur in card games, when players publicly show cards to some others, but not to all. But card updates can also blow up the size of a model. Suppose I opened my envelope, but you cannot tell if I read the card in it or not. Let us say that in fact, I did look. In that case, the intuitive update is to the model



The road to this broader kind of update leads via Gerbrandy 1999, Baltag, Moss & Solecki 1999, van Ditmarsch 2000. The general idea is this.

Complex communication involves two ingredients: a current information model $\boldsymbol{M}$, and another epistemic model $\boldsymbol{A}$ of possible physical actions, which agents may not be able to distinguish. Moreover, these actions come with *preconditions* on their successful execution. E.g., truthful public announcement *A!* can only happen in worlds where *A* holds. General update takes a *product* of the two models, giving a new information model $\boldsymbol{M} \times \boldsymbol{A}$ whose states (s, a) record new actions taken at s, provided the preconditions of a is satisfied in $\boldsymbol{M}, s$. This may transform the old model $\boldsymbol{M}$ drastically. The basic epistemic stipulation is this. Uncertainty among new states can only come from existing uncertainty via indistinguishable actions:

$$(s, a) \sim_i (t, b) \quad \text{iff} \quad \text{both } s \sim_i t \text{ and } a \sim_i b$$

In the first card example, the actions were "read lecture", "read night out". Taking preconditions into account and computing the new uncertainties gives the correct

(L, read lecture) —*you*— *(N, read night out)*

The second example involved a third action "do nothing" with precondition True, which I can distinguish from the first two, but you cannot. Product update delivers a model with four worlds *(L, read L). (L, do nothing), (N, read N), (N, do nothing)* – with agents' uncertainties precisely as shown above.

Clearly, truth values can change drastically in product update. Dynamic-epistemic logic now gets very exciting, involving combining epistemic formulas true in $\boldsymbol{M}$ and epistemic information about $\boldsymbol{A}$ expressed in a suitable language. Many of the concerns for public update in this paper will return in more sophisticated versions.

General communication General tools like this can chart many varieties of communication, and their broad patterns. For instance, there are natural *thresholds*. One leads from partial information inherent in a game or a communicative convention to that generated by people's limitations, such as bounded memory or limited attention span. Another, perhaps more exciting, crosses from mere partial information to misleading, lying and cheating. In principle, product update also describes the latter, but there is a lot of fine-structure to be understood. A final broad challenge are hiding mechanisms, such as security protocols on the Internet.

Games and social software Again, there is not just analysis, but also synthesis. Communication involves planning what we say, for a purpose. The proper broader setting for this are *games*, which involve preferences and strategies (cf. Baltag 2001, van Benthem 1999–2002). This is one instance of what has been called 'social software' recently: the design of mechanisms satisfying epistemic specifications such as who gets to know what (cf. Parikh 2002, Pauly 2001).

Communication channels After all these sweeping vistas, we come down to earth with a simple example, again based on a puzzle. The 1998 'National Science Quiz' of the Dutch national research agency NWO had the following question:

> Six people each know a secret. In one telephone call, two of them can share all secrets they have. What is the minimal number of calls they have to make to ensure that all secrets become known to everyone?

The answers offered were: *7, 8, 9*. The correct one turns out to 8. For *N* people, *2N-4* calls turns out to be optimal, a result which is not deep but difficult to prove. The best algorithm notes that four people can share all their secrets in four steps:

> *1 calls 2, 3 calls 4, 1 calls 3, 2 calls 4.*

So, single out any four people in the total group.

> First let the other *N-4* call one of them, then let the four people share all they have, then let the *N-4* people call back to their informant.

The total number of calls will be *2N-4*. Now, this clearly raises a general question. What happens to update logic when we make a further semantic parameter explicit, viz. the *communication network*? Our running example of public announcement presupposed some public broadcast system. The gossip puzzle assumes two-by-two telephone connections without conference calls. We can look for results linking up desired outcomes with properties of the network. E.g., it is easy to show that

> Universal knowledge of secrets can be achieved if and only if the network is *connected*: every two people must be connectible by some sequence of calls.

But there are many other intriguing phenomena. Suppose three generals with armies on different hilltops are planning a joint attack on the adversary in the plain below. They have completely reliable two-way telephone lines. One of them possesses some piece of information p which has to become common knowledge among them in order to execute a successful coordinated attack. Can they achieve this common knowledge of p? The answer is that it depends on the scenario.

If the generals only communicate secrets, even including information about all calls they made, then common knowledge is unattainable, just as in the more familiar two-generals problem with unreliable communication. Informally, there is always someone who is not sure that the last communication took place. More precisely, product update allowing for this uncertainty will leave at least one agent uncertainty chain from the actual world to a $\neg p$–world, preventing common knowledge. But what about general A phoning general B, sharing the information, and telling him that he will call C, tell him about this whole conversation, including the promise to call him? This is like mediators going back and forth between estranged parties. Can this produce common knowledge? Again, it depends on whether agents are sure that promises are carried out. If they are, then a scenario arises with actions and observations where product update will indeed deliver common knowledge.

We leave matters at this informal state here. Our aim in this excursion has merely been to show that update logic fits naturally with other aspects of communication, such as the availablitiy and reliability of channels.

10 Logic and communication

Traditionally, logic is about reasoning. If I want to find out something, I sit back in my chair, close my eyes, and think. Of course, I might also just go out, and ask someone, but this seems like cheating. At the University of Groningen, we once did a seminar reading Newton's two great works: *Principia Mathematica*, and the *Optics*. The first was fine: pure deduction, and facts only admitted when they do not spoil the show. But the *Optics* was a shock, for being so terribly unprincipled!

Its essential axioms even include some brute facts, for which you have to go out on a sunny day, and see what light does on when falling on prisms or films. For Newton, what we can observe is as hard a fact as what we can deduce. The same is true for ordinary life: questions to nature, or other knowledgeable sources such as people, provide hard information. And the general point of this paper is that logic has a lot to say about this, too. One can see this as an extension of the agenda, and it certainly is. But eventually, it may also have repercussions for the original heartland. Say, what would be the crucial desirable meta-properties of first-order logic when we add the analysis of communication as one of its core tasks?

I will not elaborate on this, as this paper has already taken up too much of the reader's time. And in any case, now that I know the road to the Colosseum, it is time for me to go. As a former member of the guard, I have some duties to fulfil, even though the script of the movie does not promise a happy ending.

11 References

H. Andréka, J. van Benthem & I. Németi, 1998, 'Modal Logics and Bounded Fragments of Predicate Logic', *Journal of Philosophical Logic* 27:3, 217–274.

A. Baltag, 2001, 'Logics for Insecure Communication', *Proceedings TARK VIII*, Morgan Kaufmann, Los Altos, 111–121.

A. Baltag, L. Moss & S. Solecki, 1998, 'The Logic of Public Announcements, Common Knowledge and Private Suspicions, *Proceedings TARK 1998*, 43–56, Morgan Kaufmann Publishers, Los Altos.

A. Baltag, L. Moss & S. Solecki, 2002, Updated version of the preceding, department of cognitive science, Indiana University, Bloomington, and department of computing, Oxford University.

J. Barwise & L. Moss, 1997, *Vicious Circles*, CSLI Publications, Stanford.

J. van Benthem, 1996, *Exploring Logical Dynamics*, CSLI Publications, Stanford.

J. van Benthem, 1997, 'Dynamic Bits and Pieces', Report LP-97-01, Institute for Logic, Language and Computation, University of Amsterdam.

J. van Benthem, 1998, 'Dynamic Odds and Ends', Report ML-98-08, Institute for Logic, Language and Computation, University of Amsterdam.

J. van Benthem, 1999A, 'Radical Epistemic Dynamic Logic', note for course 'Logic in Games', Institute for Logic, Language and Computation, University of Amsterdam.

J. van Benthem, 1999B, 'Update as Relativization', manuscript, Institute for Logic, Language and Computation, University of Amsterdam.

J. van Benthem, 1999-2002, *Logic in Games*, lecture notes, Institute for Logic, Language and Computation, University of Amsterdam, http://staff.science.uva.nl/~johan

J. van Benthem, 2000, 'Update Delights', invited lecture, ESSLLI Summer School, Birmingham, and manuscript, Institute for Logic, Language and Computation, University of Amsterdam.

J. van Benthem, 2001A, 'Games in Dynamic Epistemic Logic', in G. Bonanno & W. van der Hoek, eds., *Bulletin of Economic Research* 53:4, 219–248.

J. van Benthem, 2001B, 'Logics for Information Update', *Proceedings TARK VIII*, Morgan Kaufmann, Los Altos, 51–88.

J. van Benthem, 2001C, 'Structural Properties of Dynamic Reasoning', invited lecture, *Dynamics 2001*, Prague. To appear in J. Peregrin, ed., *Meaning and the Dynamic Turn,* Elsevier Science Publishers, Amsterdam.

J. van Benthem, 2002A, private correspondence with John McCarthy and Alexandru Baltag.

J. van Benthem, 2002B, 'Rational Dynamics', invited lecture, LOGAMAS workshop, department of computer science, University of Liverpool. Manuscript, Institute for Logic, Language and Computation, Amsterdam.

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

H. van Ditmarsch, 2000, *Knowledge Games*, dissertation DS-2000-06, Institute for Logic, Language and Computation, University of Amsterdam.

H. van Ditmarsch, 2002, 'Keeping Secrets with Public Communication', department of computer science, University of Otago.

H-D Ebbinghaus & J. Flum, 1995, *Finite Model Theory*, Springer, Berlin.

R. Fagin, J. Halpern, Y. Moses & M. Vardi, 1995, *Reasoning about Knowledge*, MIT Press, Cambridge (Mass.).

J. Gerbrandy, 1999, *Bisimulations on Planet Kripke*, dissertation DS-1999-01, Institute for Logic, Language and Computation, University of Amsterdam.

J. Halpern & M. Vardi, 1989, ‘The Complexity of Reasoning about Knowledge and Time’, *Journal of Computer and Systems Science* 38:1,195-237.

M. Hollenberg, 1998, *Logic and Bisimulation*, dissertation, Publications Zeno Institute of Philosophy, vol. XIV, University of Utrecht.

D. Kozen, D. Harel & J. Tiuryn, 2000, *Dynamic Logic*, MIT Press, Cambridge (Mass.).

J. McCarthy, 2002, 'Two Puzzles about Knowledge', department of computer science, Stanford University, http://www.stanford.edu/~jmc

M. Marx & Y. Venema, 1997, *Multi-Dimensional Modal Logic*, Kluwer Academic Publishers, Dordrecht.

B. Moore, 1985, 'A Formal Theory of Knowledge and Action', research report, SRI International, Menlo Park.

Nationale Wetenschapsquiz 1998, 'Gossip Puzzle', http://www.nwo.nl/

R. Parikh, 2002, 'Social Software', *Synthese* 132, 187–211.

M. Pauly, 2001, *Logic for Social Software*, dissertation DS-2001-10, Institute for Logic, Language and Computation, University of Amsterdam.

J. Plaza, 1989, 'Logics of Public Announcements', *Proceedings 4th International Symposium on Methodologies for Intelligent Systems*.

Spinoza brochure *Logic in Action*, 1998, Institute for Logic, Language and Computation, University of Amsterdam.

C. Stirling, 1999, 'Bisimulation, Modal Logic, and Model Checking Games, *Logic Journal of the IGPL 7:1*, 103–124. Special issue on Temporal Logic, edited by A. Montanari & Y. Venema.

M. Wooldridge, 2002, *An Introduction to Multi-Agent Systems*, John Wiley, Colchester.