Logic, Rational Agency, and Intelligent Interaction

Johan van Benthem
University of Amsterdam & Stanford University,
http://staff.science.uva.nl/~johan/

Abstract This paper records a lecture given for many audiences, including the *Workshop on Knowledge and Rationality*, 18th European Summer School in Logic, Language and Information, Malaga, August 2005, the *Dynamic Logic Workshop* in Montréal, June 2007, and the Philosophical Logic Section, *DLMPS Congress* Beijing, August 2007. We make a plea for recasting logic as a theory of interactive agency, and show how this perspective fits both old achievements and new broader ambitions for the field.

1 Logic as information flow: a new view, but also an old one

The restaurant: entangled informational processes When asked at public occasions to explain what logic is, I often use the following evergreen scenario, showing our discipline at work every day in our city. You are in a café with two friends, and the three of you have ordered a beer, a wine, and a water. Now some new person comes back with three glasses. What will happen? Everyone agrees that three things occur in sequence:

> First the waiter asks "Who has the wine?", say, and puts that glass.
> Then, he asks who has the beer, and puts that glass. And then, he
> does not ask any more, but just puts the remaining glass.
> Two questions, and then one inference!

When he puts that third glass without asking, you observe a logical inference in action: the information in the two answers received allows the waiter to deduce where the third one must go. [1] One can spell out this final stage in terms of a valid propositional schema

$$A \vee B \vee C, \neg A, \neg B \Box\!\!- C,$$

but this is not my main point here. To me, this old example cries out for a new twist. There is a natural unity to this scenario. The waiter first obtains the relevant information by communication and perhaps observation, and then, once enough data have accumulated, he infers an explicit solution. Now on the traditional line, only the latter deductive step is the proper domain of logic, while the former are at best 'pragmatics'. But in my view,

[1] When hearing this example, the President of Amsterdam University gave me a warning: "Johan, you should be more careful and avoid low-class cafés. When *I* order something, I am not paying all that money to just have my glass put in front of me, while the others also get a question." Indeed, good waiters put the last glass with a smile, or they will even say: "So this must be you". The calculus of politeness has its own laws on top of logic.

both informational processes are on a par, and both should be within the compass of logic, which is about information flow in general, not just deductive elucidation. In my book, asking a question and processing an answer is just as 'logical' an activity as drawing an inference! And accordingly, logical systems should be able to account for both, as observation, communication, and inference occur entangled in most meaningful activities. But what is involved in this ambitious task?

The logical challenge: 'social dynamics' First, there is the information flow itself in the café scenario. The initial part is a sequence of *update actions* on information states, viewed as sets of live options at the current stage. Initially, there are 6 ways in which three glasses can be distributed over three people. The first answer reduces this uncertainty from 6 to 2, and the second answer reduces it to 1, i.e., the actual situation:

(6) *answer 1>* (2) *answer 2>* (1)

That is why no third question is needed: one just spells out the situation. The *dynamics of informational actions* is an obvious target for logical theory. And to make this work, we must also give an account of the underlying statics: the information states that the actions work over. I will argue that this can be done, though much remains to be understood. [2] But there is another striking feature to the informational activity going on in the restaurant. Questions and answers typically involve more than one agent, and hence the dynamics is *social*, having to do with what people know about each other. In particular, the waiter asks us, because he knows that we know what we ordered. This higher-order knowledge about others is crucial to human communication and interaction in general.

A historical pedigree after all Is all this merely new-fangled tinkering with the good old core values of logic? I do not think so. The ideas put forward here are themselves ancient and obvious. For instance, traditional Indian logic distinguished three principled ways of getting information. The easiest route is to observe, when that is possible. The next method is inference, in case observation is impossible or dangerous, as with a coiled object in a room where we cannot see whether it is a piece of rope, or a cobra. And if these two methods fail, we can still resort to communication, and ask some expert. Similar ideas occur in medieval Western logic, and the restaurant scenario shows that the same natural combination occur today. [3] Moreover, the social interactive aspect of information flow is

---

[2] For instance, the waiter's final act of *inference* also produces 'new information' – but surely, not the same kind as that produced by the initial observational updates. An explicit model for this additional inferential information would have to employ syntactic fine-structure (cf. the survey of information paradigms inside logic in van Benthem & Martinez 2007), but we will only touch on this issue lightly in Sections 8 and 11 below.

[3] At a 2005 Winter School at IIT Bombay, Mumbai (cf. Gupta, Parikh & van Benthem, eds., 2007), I once presented this point to students with the question whether there was

just as ancient, going back to the very roots of logic. While many people see Euclid's *Elements* as the paradigm for logic, with its crystalline structure of mathematical proofs and eternal insights, the true origin of the discipline may be closer to Plato's *Dialogues*, an argumentative practice with clear patterns of confirmation and refutation between participants It has been claimed that logic arose originally out of political and legal debate in all its three main traditions: Chinese, Indian, and Western. [4] And this multi-agent interactive view has emerged anew in modern times. A beautiful case are the so-called *dialogue games* from the mid 1950s (Lorenzen 1955), which explained logical validity in terms of winning strategies for a proponent arguing the conclusion against an opponent granting the premises.

So, it seems that logical activity is interactive, and that its theory should reflect this. Once again, some colleagues find this alarming, as social multi-agent aspects are dangerously reminiscent of gossip, status, and Sartre's famous warning that "Hell is the Others". Maybe the best way of dispelling such fears is taking a look at what all this entails:

2 Information flow in dynamic epistemic logic

Questions and answers Take the striking multi-agent phenomenon of communication. A lot of interesting logical structure already shows in very simple question/answer examples, the ubiquitous building blocks of interaction. Consider the following dialogue:

Me: "Is this the Forbidden City? "
You: "No."
You: "It is the Friendship Hotel."

What this certainly conveys are facts about the current location. But there is much more going on. By asking the question, at least in a normal scenario (not, say, a competitive game), I indicate that I do not know the answer. And by asking you, I also indicate that I think that you may know the answer, again under normal circumstances. [5] Moreover, your

---

beer on campus. I had tried observation, inspecting most buildings for alcohol outlets the night before. I had tried deduction, reading all the conference material through and through. And now I was reduced to asking experts, viz. the students. No answer was forthcoming right then, but in the evening two students arrived carrying a plastic bag. What they said was this: "Sir, the answer to your question is 'No'. However, there is a liquor store right outside the campus gate, and since we thought you were asking the question because you needed beer, we bought you three bottles."

[4] E.g., the Mohist literature in early China discusses the Law of Non-Contradiction as an interactive principle of rational conversation: 'resolve contradictions with others', 'avoid contradicting yourself'. Cf. Liu & Zhang 2007.

[5] All such presuppositions are off in a classroom with a teacher questioning students.

answer and the follow-up statement do not just transfer the bare facts to me. They also make sure that you know that I know, that I know that you know that I know, and in the limit of epistemic iterations like this, they achieve so-called *common knowledge* of the relevant facts in the group consisting of you and me. This common knowledge is not a by-product of the fact transfer. It rather forms the basis of our mutual expectations about future behaviour. Thus, keeping track of 'higher-order' information about others is crucial in many disciplines, from philosophy (interactive epistemology) and linguistics (communicative paradigms of meaning) to computer science (multi-agent systems) and cognitive psychology ('theory of mind'). Indeed, the ability to move through an informational space keeping track of what other participants do and do not know, including the crucial ability to switch and view things from other people's perspective, seems characteristic of human intelligence.

While this social dynamics sounds forbidding, very simple logical systems exist which shed some light on it. We present one here, just to show how the ambitions in Section 1 can be realized without abandoning logical systems as we know them today. This will also quickly give us some further material for more concrete discussion of relevant issues.

Epistemic logic dynamified Let us quickly review the static base of our dynamic logic-to-be. To model the basics of the preceding question-answer scenario (and much more), we can use models for *epistemic logic*, proposed by Hintikka in the 1960s (Hintikka 1963) as a way of analyzing the philosopher's conception of knowledge. In what follows, however, we will read the knowledge found in this system in a modern spirit as what is true 'to the best of agents' information', with that information viewed as ranges of worlds that are still candidates for the actual situation. (Cf. van Benthem & Martinez 2007 for much further background on this move.) The language has a classical propositional base with modal operators $K_i\varphi$ ('*i knows that* φ') and $C_G\varphi$ ('φ is *common knowledge* in group G'):

$$p \mid \neg\varphi \mid \varphi \wedge \psi \mid K_i\varphi \mid C_G\varphi$$

We write $\langle i\rangle\varphi$ for the dual modality $\neg K_i\neg\varphi$: 'agent i considers φ possible'. The dual of $C_G\varphi$ is $\langle C_G\rangle\varphi$. Models $\boldsymbol{M}$ are triples $(W, \{\sim_i \mid i \in G\}, V)$, with W a set of worlds, the $\sim_i$ are binary accessibility relations between worlds, and V is a propositional valuation. [6] Then the standard epistemic truth conditions are as follows:

$\boldsymbol{M}, s \models K_i\varphi$ iff for all t with $s \sim_i t$: $\boldsymbol{M}, t \models \varphi$

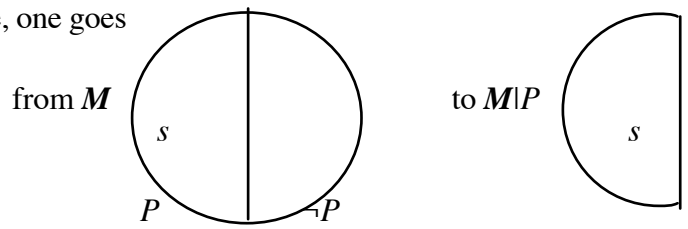
$\boldsymbol{M}, s \models C_G\varphi$ iff for all t that are reachable from s by some finite sequence of $\sim_i$ steps ($i \in G$): $\boldsymbol{M}, t \models \varphi$

[6] One often takes these relations to be equivalence relations, though this is optional.

Now comes the common sense view of information flow: incoming new information eliminates possibilities from the current range. In particular, public announcements of true propositions *P* provide 'hard information', which changes the current model as follows:

> For any model ***M***, world *s*, and formula *P* true at *s*,
> *(**M**|P, s)* (***M** relativized to P at s*) is the sub-model
> of ***M*** whose domain is the set $\{t \in \mathbf{M} \mid \mathbf{M}, t \models P\}$.

In a picture, one goes



Crucially, truth values of formulas may change in such an update step: most notably, because agents who did not know that *P* now do after the announcement.

Public announcement logic Now we must bring the dynamics of these successive update steps into our static epistemic logic. Here is how (cf. van Benthem 2006B for a survey of the following system and its technical properties). The language of *public announcement logic PAL* is the epistemic language with added action expressions:

Formulas	*P*:	$p \mid \neg\varphi \mid \varphi\ \psi \mid K_i\varphi \mid C_G\varphi \mid [A]\varphi$
Action expressions	*A*:	*!P*

The semantic clause for the dynamic action modality is as follows:

$\mathbf{M}, s \models [!P]\varphi$ iff *if* $\mathbf{M}, s \models P$, *then* $\mathbf{M}|P, s \models \varphi$

In particular, this language allows us to make typical assertions like

$[!P]K_i\varphi$ #

which state what agent *i* knows after having received the hard information that *P*. Such formulas neatly high-light the combination of ideas from diverse fields involved here. Speech acts *!P* come from linguistics and philosophy, knowledge assertions $K_i\varphi$ from philosophical logic, computer science, and economics. And the dynamic effect modality [] combining these actions and assertions comes from program logic in computer science. [7]

Reasoning about information flow in update steps like this revolves around the formula #. In particular, we need to analyze what it means in terms of a dynamic *'recursion equation'* telling us what the new knowledge is in terms of the old knowledge which the agent had before the update took place. Here is the relevant valid principle of update, which can be verified using the above truth clauses, and the above picture for concreteness:

---

[7] The notorious 'culture gap' between humanities and exact sciences is bridged in logic!

$[!P]K_i\varphi \quad \leftrightarrow \quad P \rightarrow K_i(P \rightarrow [!P]\varphi)$ [8]

We will discuss this key principle of information flow or conversation in more detail later. Here is how it functions in a complete calculus of public announcement:

Theorem *PAL* is axiomatized completely by the usual laws of epistemic logic over our chosen static model class [9] plus the following *reduction axioms*:

$[!P]q \quad \leftrightarrow \quad P \rightarrow q$ *for atomic facts* q

$[!P]\neg\varphi \quad \leftrightarrow \quad P \rightarrow \neg[!P]\varphi$

$[!P](\varphi\wedge\psi) \quad \leftrightarrow \quad [!P]\varphi \wedge [!P]\psi$

$[!P]K_i\varphi \quad \leftrightarrow \quad P \rightarrow K_i(P \rightarrow [!P]\varphi)$

$[!P]C_G(\varphi, \psi) \quad \leftrightarrow \quad C_G(P \rightarrow [!P]\varphi, [!P]\psi).$

Taken together, these axioms analyze the effects of new information compositionally. As a result, they reduce all assertions containing dynamic action modalities to basic epistemic statements about the initial model, which 'pre-encode' future effects of the dynamics. Both this recursive type of analysis and the pre-encoding power needed to make it work return in many areas of cognitive action, including belief revision, as we shall see below.

The theorem will suffice to show that logics dealing with information flow look just like systems that we know, and can be developed maintaining the same technical standards. [10] Often, the dynamic super-structure can be fitted onto an already existing logical system

---

[8] Here is a formal analysis for this recursion equation. Compare the models $(\boldsymbol{M}, s)$ and $(\boldsymbol{M}|P, s)$ before and after the update. [Drawing a picture with both helps.] The formula $[!P]K_i\varphi$ says that, in $\boldsymbol{M}|P$, all worlds $\sim_i$-accessible from s satisfy φ. The corresponding worlds in $\boldsymbol{M}$ are those $\sim_i$-accessible from s *which satisfy* P. Given that truth values of formulas may change in an update step, the right description of these worlds in $\boldsymbol{M}$ is not that that they satisfy φ (which they do in $\boldsymbol{M}|P$), but rather $[!P]\varphi$: they become φ after the update. Finally, $!P$ is a partial function: P must be true for its public announcement. Thus, we need to make our assertion on the right conditional on $!P$ being executable, i.e., P being true. Putting this together, $[P!]K_i\varphi$ says the same as $P \rightarrow K_i(P \rightarrow [!P]\varphi)$. The latter can be simplified to the equivalent formula $P \rightarrow K_i[!P]\varphi$ usually found in the literature.

[9] Here the binary epistemic operator $C_G(\varphi, \psi)$ of 'conditional common knowledge' has a technical function, explained in detail in van Benthem, van Eijck & Kooi 2006.

[10] Note that relating this system to specific applications involves the choice of a model and a set of proposition letters. Thus, what are the relevant 'possible worlds' is itself a process of representation, and it might even change in the course of a conversation. E.g., with my question about the Forbidden City, originally just one proposition letter plays a role, but your coda about the Friendship Hotel now makes a second one relevant, transforming the model. This additional 'dynamics of representation' is less-understood.

describing properties of the static 'snapshots' of the relevant informational process. Indeed, there is a growing literature on the model theory, proof theory, and computational complexity of public announcement logic and its more sophisticated variants.

3 Agents: the fine-structure of inference and observation

The logic of public announcement is fully capable of dealing with the information flow in our original café example. Moreover, it performs the two basic tasks involved there in tandem, describing information flow through both observations and inferences by agents. [11] Now there is a more general ambition at play here. The dynamic logics of this paper develop the notion of a *rational agent*, as a much richer counterpart to the austere and solitary paradigmatic proof systems or computational devices in traditional logic. One important theme is then what notion of agency emerges from our considerations.

Idealized agents What sort of agents populate *PAL*? For a start, epistemic logic makes some sweeping idealizations. Agents are *'omniscient'*: their knowledge is closed under all inference rules of the system, and on the usual semantics also *'introspective'*: they know when they know (and they also know when they do not know). Our dynamic analysis has nothing to say per se about these two idealizations. [12] Instead, we have added one more!

The recursive equation in the *PAL* axiom for $[!P]K_i\varphi$ embodies a further idealized ability of agents in the dynamics of sequential actions, viz. *perfect memory*. World elimination encodes what has taken place in the current state, in a wholly transparent manner clear to all. Technically, we can see this as follows. Disregarding some syntax, the *PAL* axiom essentially performs an operator switch between $[\,]K$ and $K[\,]$. And such switches encode strong assumptions on memory and observation. Consider the putative logical principle

$K[a]\varphi \rightarrow [a]K\varphi$

if I know that doing *a* will produce φ, then after doing *a,* I know that φ holds.

While this is correct for transparent publicly observable actions by agents who recall their previous state, it fails for actions which impair epistemic abilities (cf. Moore 1985). In particular, what is at stake here is memory. I know now that after drinking, I get boring.

---

[11] I use the term 'observation' for the basic ability now, since the communication in our simple examples may be viewed as the special case of observation of what the others say. Dynamic-epistemic logics are usually presented with an emphasis on communication, but I personally feel they are best understood as logics of observation in multi-agent settings.

[12] If we make the inferential process explicitly dynamic, however, with suitable syntactic information states modified by inference steps, then omniscience can be blocked. Sections 8, 11 have some references on how to do this – but there is no consensus in the literature!

But the tragedy of drinking is that, after I have drunk, I do not know that I am boring. [13] Likewise, the converse of the preceding axiom, also present in the *PAL* recursion equation, expresses a learning principle called *'no miracles'*: if I am uncertain now between two worlds, then seeing the same action in both is not going to remove my uncertainty.

Diversity and parametrized powers The upshot of our discussion is this. Information flow essentially happens to *agents* who use it in various ways. But current dynamic epistemic logics do not just give a neutral description of arbitrary participants in this process, they merely describe what idealized agents should be able to observe and infer. And this idealization raises a question. The reality of life is *diversity of agents*, with different bounds on their inferential, introspective, and observational powers, as well as different bounds on memory. Indeed, one hallmark of rational behaviour seems to be our ability to function successfully in environments with agents of very different skills and inclinations. And correspondingly, we may want our logical systems 'parametrized', so that the interplay of different agents can be accommodated smoothly.

While some attempts exist in this direction, including dynamic logics making different assumptions about memory capacity (cf. Liu 2006), there is no standard way so far of doing this all across the board. In particular, while there is some highly suggestive proof-theoretic literature on bounded agents manipulating syntactic 'evidence' (cf. Artemov 1994), we still lack a canonical way of representing inferential abilities of agents in a parametrized fashion. More generally, unlike the case of computation and Turing Machines, and the Automata Hierarchy providing the fine-structure underneath them, we still lack a universal model of how an agent works, let alone one that can be parametrized for varying abilities. In what follows, we merely explore some bits and pieces of this area. Of course, these are tantalizing bits and pieces, otherwise we might just as well stop here!

Full dynamic epistemic logic One encouraging fact is that dynamic epistemic logics do have a full-fledged account of diversity in *powers of observation* (cf. Footnote 11). Often, agents cannot fully observe a situation – witness the earlier Indian 'coiled rope, or cobra'. Moreover, different agents can have different observational access to a situation. Think of a card game, where you draw a card from the stack, but the other players do not see which one. The total effect of such mixtures can be hard to describe, witness the complications arising from using emails with lots of *cc*'s and *bcc*'s. In truthful public announcement, there is just one event, publicly visible to all, and the precondition for it to happen is that the announced proposition be true. This line of thinking can be generalized to scenarios involving many possible events, like drawing different cards from a stack, where agents

[13] Thus, our café example was a bit tricky. In any case, logical waiters should not drink!

may not be able to distinguish them in the same way (cf. Baltag, Moss & Solecki 1998). This requires dynamic structures much like epistemic models. *Event models* are structures

$$\boldsymbol{A} = (E, \{\sim_i \mid i \in G\}, \{PRE_e \mid e \in E\})$$

consisting of relevant events, and relations $\sim_i$ encoding what agents cannot distinguish. Events *e* also have *preconditions* PRE_e for their successful execution: the red card lying on top for my drawing it, your knowing the answer to my question, etc. These provide the core information when we observe the event. Now here is the general *Update Rule*: for any epistemic model $(\boldsymbol{M}, s)$ and event model $(\boldsymbol{A}, e)$, the *product model* $(\boldsymbol{MxA}, (s, e))$ has

Domain $\{(s, e) \mid s$ a world in $\boldsymbol{M}$, e an event in $\boldsymbol{A}$, $(\boldsymbol{M}, s) \models PRE_e\}$,
Accessibility $(s, e) \sim_i (t, f)$ iff *both* $s \sim_i t$ *and* $e \sim_i f$,
The valuation for atomic propositions p at (s, e) is just that at s in $\boldsymbol{M}$. [14]

Product update models a great many scenarios. It deals with misleading actions as well as truthful ones, and with *belief* as well as knowledge. In particular, the smooth course of world elimination is now much more tortuous. Epistemic models can easily get *larger* as product update proceeds, as happens in parlour games or email scenarios with *bcc*'s.

The corresponding language is again an epistemic one plus this time new action modalities containing descriptions of event models, interpreted as follows:

$$\boldsymbol{M}, s \models [\boldsymbol{A}, e]\varphi \text{ iff } \boldsymbol{M} \boldsymbol{x} \boldsymbol{A}, (s, e) \models \varphi.$$

Again, the logic for this system is effectively axiomatizable and decidable. And again, the key is finding the recursion equation for knowledge after some complex epistemic event:

$$[\boldsymbol{A}, e]K_i\varphi \quad \leftrightarrow \quad PRE_e \wedge \bigwedge\{ K_i[\boldsymbol{A}, f]\varphi) \mid f \sim_i e \text{ in } \boldsymbol{A}\}$$

There is much further literature on this system (van Benthem, van Eijck & Kooi 2006, van Ditmarsch, van der Hoek & kooi 2007, Baltag, van Ditmarsch & Moss 2007), including such topics as addition of common knowledge and other forms of group knowledge, extensions to richer modal fixed-point languages, a range of inter-disciplinary connections to security and process algebra, epistemic temporal logic, game theory, and so on.

4 Warm-blooded agents: revision and learning

Agents who correctly record all information from their observations, and industriously draw the right conclusions from their evidence, may be rational in some Olympian sense. At the same time, they are cold-blooded recording devices. But rationality does not reside in always being cautious, and always being right. It can be argued, with Karl Popper, that its peak moments only occur with 'warm-blooded agents', who are opinionated, make

[14] This stipulation can be generalized to deal with genuine changes in the world.

mistakes, and then: *correct* themselves. [15] Thus, rationality is also about the dynamics of revision and learning. In a very concrete setting, revision comes to the fore in conversation, our original example. People contradict each other, and then something more spectacular has to happen than mere update. Maybe one of them was wrong, maybe they all were, and they have to adjust. Modeling this involves a further distinction between information coming from some source, and agents' various attitudes and responses to it. Standard references for 'belief revision theory' are Gärdenfors & Rott 1995, Rott 2007, while in what follows here, we mainly take the dynamic logic-based line of van Benthem 2007A.

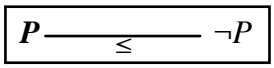
Belief and plausibility order Agents can have other attitudes toward propositions than our earlier 'knowledge', in particular, *beliefs* that can turn out incorrect. Standard logics of belief analyze assertions $B_i\varphi$ for '*agent i believes that* φ'. Their semantics adds a new idea to the information ranges in our epistemic modeling so far. We assume further gradations, in the form of a *plausibility ordering* of worlds as seen from some vantage point:

$\leq_{i,s} xy$ in world s, agent i considers y at least as plausible as x.

Thus, while the earlier ranges of epistemic alternatives corresponded to the strict information that we have, the same ranges ordered by plausibility give finer gradations. In particular, we now define belief semantically as '*truth in the most plausible options*':

$\boldsymbol{M}, s \models B_i\varphi$ iff $\boldsymbol{M}, t \models \varphi$ for all t which are maximal in the ordering $\varphi xy. \leq_{i,s} xy$.

Here is an elementary example. Consider a model with two worlds that are epistemically accessible, but the one with ¬P considered more plausible than the other:



At the actual world with ***P***, the agent does not know whether *P*, but she does (mistakenly!) believe that ¬*P*. It is crucial that our beliefs can be false. [16] As with epistemic logic, there are complete doxastic logics and a whole theory around them (cf. Fagin et al. 1995). [17]

Next, in doxastic logic, one soon finds that mere beliefs are not sufficient for explaining agents' behaviour. We want to know what they would believe were they to receive new information. This *pre-encoding*, in our earlier sense, requires *conditional belief*:

---

[15] Compare a lecture with a mathematician writing a proof on a blackboard to a research colloquium with people guessing, spotting problems, and making brilliant recoveries…

[16] There are some complications making this work in infinite models, but this is the idea.

[17] Most logics also analyze the *interplay* between knowledge and belief in information models with two relations $\sim_i, \leq_j$ entangled in various ways, reflecting a stand on whether knowledge implies belief, or whether one knows one's beliefs. While relations between attitudes toward information are an important topic, we focus on belief in what follows.

M, $s \models B_i^{\varphi}\psi$ iff ***M***, $t \models \psi$ for all worlds t which are *maximal* for φ in $\{x \mid x \leq_{i,s} y \text{ for some } y \in \{u \mid \mathbf{M}, u \models \varphi\}\}$.

Conditional beliefs $B_i^{\varphi}\psi$ are like logical conditionals in general (cf. Lewis 1973), in that they express what might happen under different circumstances from where we are now. [18]

Changing beliefs under hard and soft information Combined logics of knowledge and belief suggest a richer picture than what we had so far. There is *hard information* encoded in the current range of epistemically accessible worlds. But these ranges also carry fine-structure, through plausibility orderings. The latter can be viewed as the result of receiving *soft information* about some proposition *P,* making it more plausible, but not necessarily ruling out ¬*P*-worlds. Now the dynamic perspective on information change of *PAL* and *DEL* also applies to our beliefs, and how to revise these on the basis of incoming information. This process involves changes, not in the range of available worlds or epistemic accessibility, but rather in the *plausibility orderings* $\leq_{i,s}$ among worlds.

First, when we receive hard information *!P*, update proceeds by world elimination as before. We then get new beliefs related to our earlier conditional beliefs, – indeed, new conditional beliefs – and the crucial recursion equation driving the logic will say just how:

Theorem The logic of conditional belief under public announcements is axiomatized completely by (a) any complete base logic of $B_i^{\varphi}\psi$ for one's chosen models, (b) *PAL* reduction axioms, plus (c) a reduction axiom for conditional beliefs:
$[!P]\, B_i^{\varphi}\psi \leftrightarrow \ldots \quad P \rightarrow B_i^{P \wedge [!P]\varphi}[!P]\psi$

Hard information already involves non-trivial phenomena. For instance, true information can be misleading to rational agents! Consider this model with actual world 1, where all worlds are epistemically accessible, with a plausibility ordering $1 \leq 2 \leq 3$. Here the agent believes that *p* in 1, but for the wrong reason, as she considers world 3 most plausible:

1 *p, q* ≤ 2 *r* ≤ 3 *p, s*

Now suppose that a true public announcement $!\neg s$ is made. This eliminates world *3*, but in the remaining model with domain *{1, 2}:*

1 *p, q* ≤ 2 *r*

in the actual world, the agent has now come to believe, incorrectly, that ¬*p*!

Following earlier ideas of Stalnaker, Baltag & Smets 2006 have emphasized that there is another natural attitude here, viz. of *safe beliefs* which cannot be changed by new true

[18] The analogy is so close that conditional belief on reflexive transitive plausibility models satisfies exactly the laws of the minimal conditional logic (cf. Veltman 1985).

information. This provides an additional robustness, moving them closer to knowledge. Technically, safe beliefs are about those formulas φ which hold *in all worlds that are least as plausible as the current one*. Thus, a dynamic perspective on information change can also suggest new static epistemic-doxastic operators.

Genuine belief revision Next, consider the 'seasoned learner' receiving *soft information* concerning a proposition *P*. This just increases her 'preference' for *P*-worlds, without totally ruling out the others. Soft information leads to plausibility change, not world elimination. This can come in various sorts, reflecting another source of diversity for agents. A quite typical 'belief revision policy' is *lexicographic upgrade* $\Uparrow P$ (Segerberg 1995) which replaces the current ordering relation ≤ between worlds by the following:

> *all P-worlds become better than all ¬P-worlds, and*
> *within those two zones, the old ordering remains.*

Belief changes under such policies can be axiomatized completely. E.g., the logic for $\Uparrow P$ revision is found in van Benthem 2007A. Its key recursion equation reads as follows:

$$[\Uparrow P]\, B\varphi \;\leftrightarrow\; ((\langle\rangle(P \wedge [\Uparrow P]\varphi) \wedge B^{P \wedge [\Uparrow P]\varphi}\, [\Uparrow P]\varphi) \vee (\neg(\langle\rangle(P \wedge [\Uparrow P]\varphi) \wedge B^{[\Uparrow P]\varphi}\, [\Uparrow P]\varphi))$$ [19]

This formula looks complex, and we will not give a detailed explanation here. But after all, we are describing a more subtle informational process now than mere epistemic update.

Richer dynamic doxastic logics handle many further policies, such as softer variants placing just the *most plausible P*-worlds on top, leaving all others in their old position. Again, a move can be made then similar to the one from *PAL* to *DEL* in Section 2. Instead of charting different policies one by one, we can also 'enrich the trigger' for revision, making information come in the form of event models of signals to which agents assign different plausibilities. Cf. Baltag & Smets 2006, Baltag, van Ditmarsch & Moss 2007 for the resulting view of belief revision via one 'Priority Rule' applied to a variety of inputs.

Toward real learning Our main point is that belief revision for self-correcting agents is moving within the scope of standard logical theory. While satisfying, this story is far from complete. Genuine *learning* is not just single revision steps triggered by a current signal. It is about longer-term methods producing responses to input streams over time, and these are studied in formal Learning Theory (Kelly 1996). A merge of this research area with that of dynamic logic still needs to be made (cf. Hendricks 2003 for some first attempts).

---

[19] Here <> is the existential epistemic modality 'in some world of the current range'.

5 Interaction over time and games

Temporal perspective also comes in with the next crucial feature of rational agents, their engaging in purposeful behaviour over time, responding to, and influencing others. Even the simplest form of conversation, our original example, involves saying more things than one and choosing assertions depending on what others say. This interactive aspect extends the scope of our dynamic logics in various ways.

Program structures Conversation has crucial timing aspects. We say things in a certain order, what we say may depend on circumstances, and we may have to keep repeating assertions until some intended effect obtains, as in flattery or threats. Conversational plans or programs use three well-known operations of computer programs:

(a) *sequential composition* ;
(b) *guarded choice* *IF ... THEN... ELSE....*
(c) *guarded iterations* *WHILE... DO...*

And if we allow participants to speak simultaneously, not just in turn, then conversational scenarios will even involve forms of *parallel program composition*. Now *PAL* does have some relevant validities here. For instance, its equivalence

$$[!P][!Q]\varphi \quad \leftrightarrow \quad [!(P \wedge [!P]Q)]\varphi$$

appears to say that one need not say more things than one, as a single assertion will do. But this will no longer be true when we consider arbitrary iterations.

This richer dynamic logic of conversation over time has a syntax and semantic resembling that of propositional dynamic logic *PDL* – though it is also still like *PAL* in crucial ways. [20] But nevertheless, there is a surprise when putting these two decidable systems together. The longer-term logic of potentially unbounded conversation crosses a computational threshold in terms of the complexity of validity (Miller & Moss 2005):

Theorem *PAL* with common knowledge and all *PDL* program operations added to the action part of the language is undecidable, and even non-axiomatizable.

The reason is that models for this language now have two dimensions, one 'forward' in time (where the Kleene star iteration acts as an unbounded future modality), and one 'sideways' in information models (where common knowledge provised unlimited access – and together this allows us to encode high-complexity Tiling Problems. [21] [22]

---

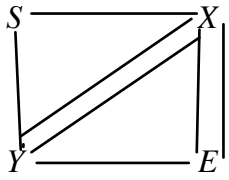
[20] For instance, all formulas are still invariant for epistemic bisimulation.

[21] Technically, validity for *PAL* with Kleene star added is even Π^1_1–complete.

[22] While it was gratifying to see that dynamic epistemic logic seemed a smash hit in Beijing in the summer of 2007, given the many posters for a pop singer called 'Update Jane', closer inspection by professional logicians led to the worrisome discovery that her

But even with this complexity, programs are still largely single-agent affairs, like a set of notes for a speech, where the audience is a largely passive recipient of your eloquence. In reality, both conversation and more general processes of learning involve a back and forth type of interaction, which is typically found in *games*.

True interaction and games To sample the spirit, consider the following not unrealistic game played between a Student and a Teacher. The Student is located at position *S* in the following diagram, but wants to reach the position of escape *E* below, whereas the Teacher wants to prevent him from getting there. Each line segment is a path that can be traveled. At each round of the game, the Teacher cuts one connection anywhere in the diagram, while Student can, and must travel one link still open to him at his current position:



If Teacher is greedy, and starts by cutting a link *S–X* or *S–Y* right in front of the Student, then it is easy to see that Student can reach *E*. However, teacher does have a *winning strategy* for preventing the Student from reaching *E*, by first cutting one line between *X* and *E*, and then letting his cutting be guided by where Student goes subsequently. Here *strategies* for players are rules telling them what to do in every eventuality. Solving games like this can be complex, emphasizing the non-trivial nature of interaction. [23] [24]

With scenarios like this, we are close to the origins of modern game theory (Osborne & Rubinstein 1994). *Zermelo's Theorem* says that extensive two-player games of finite depth with perfect information and zero-sum outcomes are *determined*: one of the two players has a winning strategy. And this result is still quite close to logic. Let the universal quantifier $[]$ range over all moves in the game, as chosen by the beginning player *I*, while the existential quantifier $\langle\rangle$ stands for moves by the other player *II*. With game depth *n*, the logical *law of excluded middle* then states the following disjunction

$$[][]\ldots[][]\text{(k times) 'player II wins'} \quad \langle\rangle\langle\rangle\ldots\langle\rangle\langle\rangle\text{ (k times)}\neg\text{ 'player II wins'}$$

---

name was really spelled 'Update * Jane', pointing at a high iterative complexity.

[23] Rohde 2005 shows that solving 'sabotage graph games' like this is *Pspace*-complete.

[24] The reader will get an even better feel for the difficulty of interaction by considering a variant. This time, the Teacher wants to force the Student to *end up in E* without any possibility of escape. Who of the two has the winning strategy this time?

and given the equivalence of *II*'s not winning and *I*'s winning, this is the determinacy. [25]

Of even more interest here is not the existence of winning strategies per se, but the fact that interacting agents may have explicit plans and strategies for dealing with each other. Such strategies have strong connections with logical notions (van Benthem 2007B), and they are an important ingredient of rational agency, also in the philosophy of action.

Infinity and temporal logic While the above games are finite, as is true of many rational activities, populations of rational agents also engage in potentially infinite processes, such as 'language use' or 'reproduction'. Describing long-term behaviour of agents over time then requires statements about these histories which may go on forever. Here the dynamic epistemic logics discussed above meet with *temporal logics* that describe properties of histories, which also have versions with knowledge and belief (cf. the discussion with extensive references in van Benthem & Pacuit 2006, van Benthem, Gerbrandy & Pacuit 2007). Such temporal systems have existed for a long time in philosophical logic and in computer science. This infinite arena is also the natural habitat of Learning Theory and even more complex structures in Game Theory are 'type spaces' for extensive games (Osborne & Rubinstein 1994). And finally, this temporal setting is also the realm of Dynamical Systems Theory, the mathematics for biology, evolutionary game theory, and even neural nets in brain research. In other words, logic in its dynamic guise lives in a larger mathematical landscape where many exciting confluences are still to be expected.

6 Preferences, goals, games, and social choice

Why questions Having gone all the way to infinity, let us now return to the basics of interaction. The Indian students of Footnote 3 did not just answer my question whether there was beer at IIT Bombay. They also asked themselves *why* I had asked my question, and then responded to that. Indeed, behind every communicative interaction, there is a "Why" question, concerning agents' goals. And 'making sense' of the interaction does not just involve meaning and information, but also mutually getting clear on those goals. This brings in another level of structure familiar from decision theory and game theory.

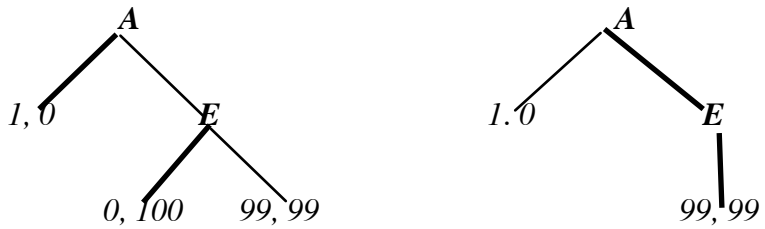
Preference logic At this stage, our dynamic logics for information flow need to add still further structure beyond knowledge and belief, viz. agents' preferences between situations, and the ways they evaluate courses of action. Right now, there is a vigorous development of logics with preference operators which try to describe preference structure (van Benthem, Girard & Roy 2007) as well as intentions (Roy 2007). Moreover, these

[25] Zermelo proved his result with a view to Chess. His result was rediscovered later by the only world champion in Chess ever produced by The Netherlands, viz. Max Euwe. For more on the 'logic games' connection behind all this, cf. van Benthem 1999.

logics come in both static and dynamic variants, with the latter describing episodes of preference change, as triggered by commands or other events with value-bestowing force (Grüne-Yanoff and Hansson, to appear; van Benthem & Liu 2007). Preference is also at the heart of social choice theory and game theory – and interesting contacts can be observed between all these areas today, with logic providing 'fine-structure'.

Logic and game theory Indeed, active interfaces between logic and game theory (de Bruin 2004, van Benthem 1999, van der Hoek & Pauly 2006) abound, and have to do with epistemic analysis of equilibrium solutions in terms of rational action, or logical analysis of games as a paradigm for interactive computation (Parikh 1985, Abramsky 2007). [26] Here we merely mention two examples showing how analysis of games naturally connects up with dynamic update logics of the sort we have discussed here. In particular, rational behaviour of players in games combines reasoning about action, belief, and preference.

Games and dynamic logics Consider the solution procedure of *Backward Induction* for extensive games, a generalization of the algorithm behind Zermelo's Theorem. Starting from outcome preferences on leaves, nodes get evaluated through the tree, representing players' intermediate beliefs as to expected outcomes and values, given that both players are acting 'rationally'. As is well-known, Backward Induction often produces 'bad equilibria' representing some socially undesirable outcome. An example is given to the left in the following picture, where the bad equilibrium *(1, 0)* predicted by reasoning about players' 'rationality' makes both hugely worse of than the cooperative outcome *(99, 99)*.



One way of doing something about this (van Benthem 2007C) is by making *promises* which change the current game through public announcements of intentions. ***E*** might promise that she will not go left, changing the game to the new one depicted on the right– and the new equilibrium *(99, 99)* results, making both players better off. Van Otterloo 2005 has a dynamic logic of players' strategic powers and preferences, where games can change by announcing more general intentions. Complete logics arise from intertwining *PAL* with modal logics of actions and preferences in a straightforward manner.

Excursion: strategies Logic of rational agency is also about strategies themselves. If we use propositional dynamic logic *PDL* to define strategies in games (van Benthem 2002),

[26] It may also be appropriate to point out that two recent recipients of the Nobel prize in Economics, viz. John Nash and Robert Aumann, have strong interests in logic as well.

adding game change leads to a joint logic *PDL+PAL* adding public announcements *[!A]*. It is easy to show that *PDL* is closed under the latter, both in its propositional and its program parts, but the crucial recursion equation now also uses an operation $\pi|A$ for programs π which basically wraps them in tests *?A*. *PDL+ PAL* is the axiomatized completely by merging the separate laws of these systems, while adding the equivalence

$$[!A]\{\pi\}\varphi \leftrightarrow (A \rightarrow \{\pi|A\}[!A]\varphi).$$

Of course, there is more structure to games than just moves and strategies, and *PAL*-style announcement scenarios also make sense with combined epistemic preference languages.

'Rational dynamics' Van Benthem 2003 uses iterated public announcement (cf. Section 5) to analyze another major arena of game theory. *Strategic games* induce epistemic models ***M*** of strategy profiles with preferences and uncertainty relations for players who know their own strategy, but not that of the others – and these come with their own solution procedures. Here a combined modal preference language can formulate statements of Weak Rationality ("no player chooses a move which she knows to be worse than some other available one") and Strong Rationality ("every player chooses a move which she thinks may be the best possible one"). When announced, these propositions eliminate worlds where they fail, and iterating these announcements to the limit, there is a smallest sub-model where *WR* or *SR* are now common knowledge. Iterated announcement of *WR* is the well-known solution concept of Iterated Removal of Strictly Dominated Strategies; and its sub-model is defined in ***M*** by a formula of a modal μ–calculus with inflationary fixed-points. [27] The same holds for iterated announcement of *SR* and game-theoretic 'Rationalizability'.

In this scenario of internal deliberation players keep recalling their 'rationality'. A similar analysis applies to extensive games. Backward Induction is obtained through repeatedly announcing that "no player chooses a move all of whose further histories end worse than all histories after some other available move". The procedure ends in largest sub-games where players have common belief of rationality. But one can also announce other types of joint agency. Van Benthem 2003 considers history-oriented versions, where players remind themselves of the *legitimate rights of others*, because of 'past favours received'.

This string of examples may suffice to illustrate the lively interface these days between logic and game theory as congenial and compatible accounts of rational agency. There is no canonical theory yet, [28] but one can definitely see contours of an interactive logic.

---

[27] If *A* has 'existential-positive' syntax (*SR* does), the definition is in standard μ–calculus.

[28] For instance, as we said earlier, dynamic logics still have to link up with perhaps the most sophisticated modeling so far, viz. 'type spaces' for games.

7 Dynamic logic and intelligent interaction generally

The claim of this paper is that logic can move beyond the standard paradigm of abstract consequence relations, or at best, a lonely theorem prover or computer to become an account of rational agents who observe, infer, communicate, learn, and interact. In a series of dynamic logics and pointers to the literature in which they lie embedded, we indicated how this might be done while sticking to the canons of logic as we know it.

This may be seen as part of a general movement. There is an emerging view today in many disciplines that theories of information, intelligence and rationality need to be *dynamic*, high-lighting human actions of communication, observation and decision, and *social*, high-lighting the interplay of several actors, often even congregating in groups. Thus, the paradigm of rational behaviour shifts from lonely thinkers writing down proofs and contemplating the ultimate truth, to the noisy realities of dialogue, debate, and colliding opinions and preferences as the locus where "la vérité s'éclate". Even within a single discipline, such as philosophy, it is striking how this interactive turn is affecting many areas in parallel, from pragmatics in the philosophy of language to interactive epistemology, and from second-person ethics to philosophy of action and social philosophy. Accordingly, we are now led to think of 'rational agents', not as solitary utility maximizers in the sense of Dickens' Scrooge – but rather as people who thrive in interactive environments, and contribute to their successful functioning.

We conclude this paper by inquiring into the status of all this, and the challenges which it poses. The logical developments presented so far look promising, with a growing body of real work. There are conferences and workshops on 'Knowledge and Rationality', journals on 'Knowledge, Rationality, and Action', and rational agency is in the air everywhere. Many of us like this 'Interactive Turn' – but what can it really achieve? I will discuss some perspectives on this perhaps surprising issue – mostly taking the earlier-mentioned developments in logic as my frame of reference. This will also serve as a way of comparing interactive logic with the traditional face of the field.

8 The actors: what is an intelligent agent?

The foundational program The preceding goals may be all to the good, but what ís a rational agent really, and what is the task we have set ourselves as theorists of intelligent interaction? One cannot consult some standard text or manifesto, because there are none.

To see the issue, think of an earlier turn which changed the face of logic, viz. the mathematization by Frege, Russell and others, and the matching hopes for the foundations of mathematics. *Hilbert's Program* provided an appealing set of goals here. The formalization would describe mathematical theories, establish their consistency, and where possible completeness, while the logic driving all this would be simple, perhaps decidable.

Thus, the goal was to clarify the methodology of exact mathematical reasoning once and for all, and scientific practice would be the better for it ever after. This was a technical enterprise with something important at stake! The great foundational discoveries of the 1930s demonstrated the infeasibility of this program, in the form of Gödel's Incompleteness Theorems, and to a lesser extent, Turing and Church's undecidability results for simple natural computational and logical problems. This course of events may be seen as a Popperian virtue by itself. At least, the foundational research program of the early 20th century had not immunized itself to criticism in a facile manner, and it had made a refutable claim. But far beyond this, the refutation had positive spin-off. Like Vergilius' Romans after the fall of Troy, logicians spread all over the academic world, and founded an empire based on positive follow-up. Just think of the Turing Machine as a universal model for computation, and recursion theory, or of insights into mathematical proofs turning into proof theory, and likewise, limits on expressive power for theories turning into model theory as the study of the variety of models allowed by logical formalisms. [29]

Toward a common model Intelligent interaction is clearly a more ambitious goal than standard computation. So, what are those actors that we have placed at centre stage in our logical enquiry? The Universal Turing Machine quickly became the general computing model, being a lucid analysis of the key features of a human computer doing sums with pencil and paper. And even though nothing similar has been found yet for the more intensional notion of an 'algorithm', this first simple rallying point focused the whole study of computation. So, suppose that our counterpart to this computational device is the Generic Rational Agent, what would be its defining skills and properties, beyond pencil and paper sums? Clearly, that analysis must be more complex than Turing's, since agents engage in such a wide array of activities, far beyond computation. [30]

I have asked many colleagues for their answer, in the form of a wish-list with features they consider constitutive of rationality. Answers were lively, but not at all conclusive. Thus, I will list some ingredients and issues which I myself find appealing and worthy of study.

---

[29] One might even say that *computer science* was born out of the debris of the foundational era, leading to a much richer agenda, including the study of intelligent agents in the by now classical *computational paradigm*. Over the years, this paradigm has come to include distributed societies of computing agents which communicate and pursue goals, whose study combines ideas from mathematical, philosophical, and computational logic. One new name for this enterprise, which borders on cognitive science, is *informatics* In this setting, our paper asks: what is the notion of agency here, and what can logic contribute?

[30] On the other hand, admittedly, computation in the modern sense has turned out to cover such diverse activities as solving puzzles, parsing natural language, or image processing.

Idealized or bounded processing powers? In line with the tradition, our dynamic logics so far idealize rational agents and endowing them with unlimited inferential and observational powers, and with ample memory to store all the fruits of all these talents. But I am also attracted by the opposite tendency in the literature, stressing the huge limitations on all these powers that human cognition operates under. In that case, the heart of rationality would be optimal performance given heart-breaking constraints. Beautiful examples of surprising optimal behaviour in that setting are found in Gigerenzer's highly original book *Simple Heuristics that Make Us Smart* (Gigerenzer 1999). This provides a 'tension': we need to explain how our logical systems can function in such a setting.

Which core tasks? And then, powers to what end? I said in the above that there is not one single core task which rational agents are called upon to perform, whereas the Turing Machine was supposed to just compute. But I also said that, under suitable encoding of data, 'computing' turned out to cover many more activities than one might think. So here is a major question. Do rational agents have a 'core business' that they must be good at?

Is it perhaps *reasoning* – as a 'normal form' reducing all other intelligent activities? I do not think so. Reasoning is indeed one important category, and we should take it in a broad sense. For instance, decision-theoretic views look 'forward' at how agents predict the future, and plan their actions. But some colleagues responding to my request for a 'core list' emphasized a 'backward-looking' talent of rational agents, viz. explaining and *rationalizing* what has already happened. Either way, next to reasoning-related tasks, other crucial abilities of rational agents such as acumen in perception and observation, and talents for successful interaction, do not reduce to 'reasoning' in any illuminating way. [31]

Revision and learning In particular, as in Section 4, I do not think that informational 'soundness': being right all the time, is a hall-mark of rational agents. To me, the peak performances of rational agents are in spotting problems, and then trying to solve them. Rationality is constant *self-correction*. This reflects my general take on the foundational collapse in the 1930s. The most interesting issue in science and mathematics is not guarantees for consistency and safe foundations, but the dynamic ability of repairing theories, and coming up with creative responses to challenges. Thus belief revision and general *learning* are the true tests of rationality in my view, rather than flawless update.

---

[31] One might say that the earlier dynamic logics do just that: reducing a wide range of rational activities to a standard formalism where validity reigns supreme. But I doubt this is the best way of thinking about what these logics achieve. And anyway, it would be a meta-level reduction of rational behaviour, rather than an object-level account of what it does.

Communication and interaction But the preceding criteria of reasoning and learning are still too restricted. Both apply to a single agent. But the core phenomenon that we are after is intelligent interaction. A truly intelligent agent can perform tasks directed toward others: ask the right questions, explain things, convince, persuade, understand strategic behaviour, synchronize beliefs and preferences with other agents, and so on. Almost paradoxically, I would state the following desideratum, which has no counterpart for Turing Machines (unless one wanted to do a Turing-style analysis for distributed computation):

A rational agent is someone who interacts rationally with other agents!

Diversity Here is another aspect of rational interaction, related to the preceding points, but worth emphasizing separately. Agents are not all the same, and they form groups whose members have diverse abilities, strategies, and so on. Understanding this diversity is a non-trivial tasks for logic (cf. our discussion of 'parametrization' in logics for agents in Section 5). Moreover, successful behaviour has to do with functioning well in a wide-range environment of agents with different capacities and habits. [32]

Switching Next to diversity, there is also a rational ability which glues us together. It is *the ability to put yourself in somebody else's place*. In its bleakest form, this is the logician's 'role switch' in a game, i.e., the interactive reading of 'negation'. But in a more concrete form, it is the ability to see social scenarios through other people's eyes, as in Kant's Categorical Imperative: "Treat others as you would wish to be treated by them."

Intelligent groups My criteria are not all exclusive, but here I list a final one. Humans typically tend to form new entities, viz. *groups*, which take on lives of their own. Indeed, our identity is made up of many layers of 'belonging' to various groups. In game theory, this leads to the study of coalitions, and on the logical side, this has led to work on common knowledge and other forms of knowledge typical for groups. [33] The earlier logical systems can help describe the fine-structure of such processes, but we are now after more global levels of rationality between individuals, groups, and even larger institutions. Thus the agenda will go all the way to analyzing, perhaps even designing, procedures for procedural justice and deliberative democracy. [34] The formation of *'rational we's'* and intelligent organizations generally needs to be acknowledged.

---

[32] It is said that Marx disliked Kropotkin, because his own narrow social range was getting on with German intellectuals like himself. By contrast, Kropotkin who came from the high Russian nobility, got on well with everyone, from intellectuals to simple workers.

[33] Maybe the most intriguing agendas are in the philosophy of action, with 'shared agency' and common intentions, and in social choice theory and judgment aggregation, which describe intelligent aggregation mechanisms for preferences and opinions.

[34] Parikh has popularized this view under the heading of 'Social Software' (Parikh 2002).

All this does not add up to one universal model of rational agency yet. The field of intelligent interaction is still waiting for its modern Turing. But I do think that these defining questions should be asked, whatever the eventual outcome – and also, that the above list stakes out a territory which needs to be covered by any eventual model.

But what is the agenda? We have staked out a set of topics concerning rational agency. We have banners like 'intelligent interaction'. We have insights and techniques from philosophy, logic, computer science, game theory, social sciences, and so on. But what is the new agenda, providing focus and unity? It is not as if we are landing on virgin shores. Like most Promised Lands in history, this one is already densely populated by many inhabitants, viz. the disciplines mentioned – to which one could add even more. Could there be an analogue to Hilbert's Program in store for us, setting a worthy goal for interactive logicians to march toward? I will not attempt to answer this intriguing question here, but rather end with some general unifying themes that run across the area.

9 Logical systems for intelligent agency: some integrating trends

Instead of having all its core definitions and long-term goals agreed on, a field might also form around a shared modus operandi. In particular, the logic of rational agency requires a process of combination. As we have shown in this paper, we have component logics for many separate tasks for agents – but in cognitive reality, these all work together. So, can we now form one grand logic of intelligent interaction? This is no simple matter.

Putting the pieces of agency together? There is not even a widely accepted model for logical agents which describes just their inferential and observational powers in tandem. This requires integrating the different notions of information that occur in logic, semantic and deductive. While there are some proposals, with joint semantic-syntactic structures that can be updated (cf. van Benthem & Martinez 2007), nothing like a consensus has emerged so far. But clearly, this is a substantial problem that must be solved.

In addition to conceptual issues, there are computational pitfalls here. The *complexity of combined logics* can be much higher than that of the components! Decidable components may create undecidable logics when the mode of combination is complex. We have seen this in the temporal epistemic logic of agents with *perfect memory*. And as for plausible modes of combination, we are far from having charted all the *conceptual entanglements* that rational agents may exhibit. [35] But again, we have to combine, and we do.

[35] Here is one: Preference is what is best for me in the worlds which I believe possible. Now an action of belief revision induces a preference change – and the reverse also happens…

Framework integration Another source of coherence are technical paradigms. Now, the area of intelligent interaction is replete with competing systems, schools, and sects. But there are some encouraging technical trends toward convergence. Gradually, contours are emerging of a common framework which might be called 'epistemic-temporal logic' in a broad sense (cf. van Benthem & Pacuit 2006). A common methodology and set of formalisms seems as powerful a source of intellectual identity as a shared language.

An interesting analogy is again with the foundational era. The 1930s saw many competing paradigms for defining computation. But eventually, it became clear that, at a well-chosen level of identification by input-output behaviour, these all described the same computable functions. *Church's Thesis* then proclaimed the unity of the field, saying all approaches described the same notion of computability – despite 'intensional differences' making one or the other more suitable for particular applications. This led to a common field of Recursion Theory, everyone got a place in the joint history, and internal sniping was replaced by external vigour. Something similar might happen in the study of intelligent interaction. If we do not have a Hilbert or Turing, we might at least have a Church.

Shared transformations A third unifying trend is the systematic transformation of traditional computational problems into broader problems of intelligent interaction. Van Benthem 2006A discusses three of these. *Epistemization* turns algorithmic tasks into versions involving information: When does a robot 'know how to stop' once in its goal region? When does an agent have the 'know-how' to achieve her goals? *Dynamization* turns static descriptions of instantaneous states of agents into processes: the earlier normative actions which induced changes in preference are a good example of a new issue brought into our scope in this way. Finally, *gamification* turns algorithmic tasks into multi-player games: the above Teacher/Student game was an example of what would normally be construed as an uneventful single-agent reachability problem. Seeing that these transformations are taking place at various places, turning computer science into something potentially much grander, might be birth pangs of a theory of rational agency.

Finally, one undeniable unifying force across the area is the *empirical reality* of intelligent interaction, and hence an independent sanity check for whatever theory we come up with.

10 Repercussions all around

To conclude this paper, let us go back to the initial line, with logics that describe a richer picture of an information-processing agent than just deductive prowess. Even at the preliminary stage sketched here, this stance has a great many repercussions.

Teaching Maybe the simplest is a didactical observation. The new logics are concrete, easily taught, and in my experience for a wide spectrum of audiences, easily understood and appreciated. If things 'fit' in teaching, there must be something to them.

Logic and information Whatever the grand issues raised in the preceding sections, the new logics also raise interesting foundational questions about logic itself. For instance, is there a unifying notion of 'information' underlying all the agent abilities we have discussed/ Van Benthem & Martinez 2007 is a sustained discussion, bringing together logical views of information as range, as correlation, and as code. The question whether these are complementary perspectives, or whether a grand unification is possible is open.

Epistemology and methodology While traditional logics seek a definition of knowledge as true belief with some further static ingredient (evidence, 'counterfactual stiffening'), our dynamic perspective provides a new approach. Knowledge is the true belief that survive interactive processes of communication and debate. This fits with views in the philosophy of science where the essence of rational enquiry is interactive, sometimes modeled by games. Thus, the opposition between formal methodology and 'sociological' views might dissolve, as there is so much formal structure to social intercourse.

Language and communication Our perspective also fits with newer views of linguistic meaning. Standard truth conditions involve no agency at all. Then 'dynamic semantics' in the 1980s introduced the single-agent idea that the meaning of a text resides in the changes it brings about in a hearer or reader. But modern game-theoretic approaches (van Rooij 2004, Gärdenfors & Warglien 2007) describe meanings as Nash equilibria in two-person communication games. While this lies below the surface of our dynamic logics, where meanings of formulas are taken to be fixed, it does fit with the general interactive analysis of key notions in logic. And there are bridges to the DEL systems of this paper. For instance, the communication games of Feinberg 2007 are about the contextual information which linguistic expressions convey over and above their literal meanings.

Social choice and groups Thinking in terms of group formation also fits naturally with logic. There is an incipient literature on logical analysis of social choice theory. But this is standard methodology, and one can be much more radical. Van Benthem 2007D analyzes belief revision itself as a process of preference aggregation, between competing signals, and characterizes the update rules of our earlier systems in terms of well-known postulates from social choice theory. Thus, what used to be single rational agents themselves turn into communities of past, present, and future signals and intentions.

Our final illustration is reserved for the heartland of the classical conception of logic.

Foundations of mathematics Well-understood, interactive multi-agent aspects have always existed right in the classical phase of logic. For instance, intuitionistic logic is about enquiring agents over time, and its analysis in Lorenzen dialogue games even made it a theory of interaction – even though this did not become part of the common understanding of constructivism. The same is true even more for linear logic as a theory

of abstract interaction, particularly in its contemporary manifestation as compositional game semantics (Abramsky 2007). [36]

Now recall Hilbert's Program, mentioned several times already. The foundational era had pathological fears of inconsistency. Frege says that, if a single contradiction were to be discovered in mathematics, "the whole building would collapse like a House of Cards". But this claim is an artefact of the wrong metaphor. Mathematics is not a house with foundations bearing the whole weight. It is rather a *planetary system* of theories with many relationships, happily spinning together in logical space. And there, contradictions are never the end of a story. To the contrary, one of the most striking ability of scientists is not to create infallible theories, but rather, having creative ways of coping with problems once they arise. [37] The history of science is replete with inventive strategies for revision. And that, of course, was the motivation for the logics for belief revision discussed as an essential part of rational agency in the above. We see human intelligence at its finest when we *correct ourselves*, learn from mistakes, and create something new out of broken dreams and refuted expectations. [38] And there is more to intelligent interaction in this arena. Mathematics is also about invention of new language, precisation when clarity is requested by a colleague, but also, of inspired informal paraphrasing when communicating the essence of a proof to one's audience. [39] Our dynamic logic should go for all of these.

---

[36] Wilfrid Sieg pointed me to a little-known passage in Turing where he, too, stresses the social aspects of learning and human intelligence as a challenge for mathematical theory.

[37] In a wonderful, little-known study from the mid 1960s (Weinberger 1965), the Czech philosopher Ota Weinberger charted persistent strategies removing inconsistencies in both common sense reasoning (disagreements in conversation) and in science. Most go back to medieval logic and beyond. One can give up assumptions, the way Zermelo-Fraenkel Set Theory traded Cantor's Full Comprehension for the Separation Axiom. Other ploys make distinctions between kinds of objects that were identified before, like 'sets' vs. 'classes' in NBG Set Theory. Another powerful strategy is the introduction of 'hidden variables', such as contextual arguments: 'I am tall for a human, but not tall for an animal.'

[38] In medical terms, the foundationalists wanted to make the whole world of mathematics free from disease, by killing off every last unclarity and inconsistency. I say, in contrast, that we should study the 'sanitary' processes of mathematical *precisation* and *revision*, which swing into action every time a new problem is discovered.

[39] In my favourite slogan these days, logic is not primarily the guardian of being right, its true role is much better described as follows: *logic is the immune system of the mind!*

13 References

S. Abramsky, 2007, 'Information, Processes and Games', Computing Lab, Oxford University. In P. Adriaans & J. van Benthem, eds., *Handbook of the Philosophy of Information*, Elsevier, Amsterdam, to appear.

S. Artemov, 1994, 'Logic of Proofs', *Annals of Pure and Applied Logic* 67, 29 – 59.

A. Baltag, H. van Ditmarsch, & L. Moss, 2007, 'Epistemic Logic and Information Update', Departemtns of computer science, Universities of Indiana, Otago, and Oxford. To appear in P. Adriaans & J. van Benthem, eds., *Handbook of the Philosophy of Information*, Elsevier Science Publishers, Amsterdam.

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 & S. Smets, 2006, 'Dynamic Belief Revision over Multi-Agent Plausibility Models', *Proceedings LOFT 2006*, Department of Computing, University of Liverpool.

J. van Benthem, 1999, *Logic in Games*, Lecture Notes, Institute for Logic, Language, and Computation, University of Amsterdam. Reprints through 2007.

J. van Benthem, 2002, 'Extensive Games as Process Models', *Journal of Logic, Language and Information* 11, 289–313.

J. van Benthem, 2003, Rational Dynamics and Epistemic Logic in Games, in S. Vannucci, ed., *Logic, Game Theory and Social Choice III*, University of Siena, Department of political economy, 19–23.Also in *International Game Theory Review* 9:1, June 2007, World Scientific, Singapore, 13 – 45.

J. van Benthem, 2006A, 'Computation as Conversation', in B. Cooper, B. Löwe & A. Sorbi, eds., to appear, *New Computational Paradigms: Changing Conceptions of What is Computable*, Springer, Heidelberg.

J. van Benthem, 2006B, 'One is a Lonely Number: on the logic of communication', in Z. Chatzidakis, P. Koepke & W. Pohlers, eds., *Logic Colloquium '02,* ASL & A.K. Peters, Wellesley MA, 96 – 129.

J. van Benthem, 2007A, 'Dynamic Logic of Belief Revision', *Journal of Applied Non-Classical Logics* 17, 129 – 155.

J. van Benthem, 2007B, 'In Praise of Strategies', to appear in J. van Eijck & R. Verbrugge, eds., *Games, Logic, and Social Software*, Report on a NIAS Project, College Publications, London.

J. van Benthem, 2007C, 'Rationalizations and Promises in Games, *Philosophical Trends*, special issue on logic, Chinese Academy of Social Sciences, Beijing.

J. van Benthem, 2007D, 'The Social Choice Behind Belief Revision', Working Paper presented at Dynamic Logic Montreal 2007, Institute for Logic, Language and Computation, University of Amsterdam.

J. van Benthem, J. van Eijck & B. Kooi, 2006, 'Logics of Communication and Change', *Information and Computation* 204(11), 1620 – 1662.

J. van Benthem, J. Gerbrandy & E. Pacuit, 2007, 'Merging Frameworks for Interaction: *DEL* and *ETL*', ILLC Amsterdam & Informatics Torino. *Proceedings TARK* 2007, University of Namur.

J. van Benthem, P. Girard & O. Roy, 2007, 'Everything Else Being Equal. A Modal Logic Approach to Ceteris Paribus Preferences', Institute for Logic, Language and Computation, University of Amsterdam.

J. van Benthem & F. Liu, 2007, 'Dynamic Logics of Preference Upgrade', *Journal of Applied Non-Classical Logics* 17, 157 – 182.

J. van Benthem & M-C. Martinez, 2007, 'The Stories of Logic and Information', Research Report, Institute for Logic, Language and Computation, University of Amsterdam. To appear in P. Adriaans & J. van Benthem, eds., *Handbook of the Philosophy of Information*, Elsevier Science Publishers, Amsterdam.

J. van Benthem & E. Pacuit, 2006, 'The Tree of Knowledge in Action', *Proceedings Advances in Modal Logic,* ANU Melbourne.

B. de Bruin, 2004, *Explaining Games*, Dissertation. ILLC, University of Amsterdam.

H. van Ditmarsch, W. van der Hoek & B. Kooi, 2007, *Dynamic Epistemic Logic*, Springer. Dordrecht.

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

Y. Feinberg, 2007, 'Meaningful Talk', in J. van Benthem, S. Ju & F. Veltman, eds., *A Meeting of the Minds*, Proceedings LORI Beijing 2007, College Publications, London, 41 – 54.

P. Gärdenfors & H. Rott, 1995, 'Belief Revision', in D. M. Gabbay, C. J. Hogger & J. A. Robinson, eds., *Handbook of Logic in Artificial Intelligence and Logic Programming* 4, Oxford University Press, Oxford 1995.

P. Gärdenfors & M. Warglien, 2007, 'Semantics, Conceptual Spaces, and the Meeting of Minds', LUCS Cognitive Science Centre, University of Lund.

G. Gigerenzer, P. Todd & ABC Group, eds., 1999, *Simple Heuristics that Make Us Smart,* Oxford University Press, New York.

T. Grüne-Yanoff & S-O Hansson, to appear, *Preference Change: Approaches from Philosophy, Economics and Psychology*, Springer, Heidelberg.

A. Gupta, R. Parikh & J. van Benthem, eds., 2007, *Logic at the Cross-Roads*, Proceedings First Indian Winter School in Logic and its Interdisciplinary Environment, IIT Mumbai 2005, Allied Publishers, Mumbai.

V. Hendricks, 2003, 'Active Agents', *Journal of Logic, Language and Information* 12, 469 – 495.

W. van der Hoek & M. Pauly, 2006, 'Modal Logic for Games and Information', in P. Blackburn, J. van Benthem & F. Wolter, eds., *Handbook of Modal Logic*, Elsevier, Amsterdam, 1077 – 1148.

D. Lewis, 1973, *Counterfactuals*, Blackwell, Oxford.

F. Liu & J. Zhang, 2007, 'Some Thoughts on Mohist Logic', in J. van Benthem, S. Ju & F. Veltman, eds., *A Meeting of the Minds*, Proceedings LORI Beijing 2007, College Publications, London, 79 – 96.

J. Hintikka, 1962, *Knowledge and Belief,* Cornel University Press, Ithaca.

K. Kelly, 1996, *The Logic of Reliable Inquiry*, Oxford University Press, New York.

F. Liu, 2006, 'Diversity of Agents', Research Report, Institute for Logic, Language and Computation, university of Amsterdam. To appear in *Journal of Logic, Language and Information.*

P. Lorenzen, 1955, *Einführung in die Operative Logik und Mathematik,* Springer, Berlin.

J. Miller & L. Moss, 2005, 'The Undecidability of Iterated Modal Relativization', *Studia Logica* 79(3): 373 – 407.

R. Moore, 1985, ' A Formal Theory of Knowledge and Action'. In J. Hobbs & R. Moore, eds., *Formal Theories of the Commonsense World*, Ablex Publishing Corp, 319-358.

M. Osborne & A. Rubinstein, 1994, *A Course in Game Theory*, The MIT Press, Cambridge (Mass.).

S. van Otterloo, 2005, *A Strategic Analysis of Multi-Agent Protocols*, Dissertation DS-2005-05, ILLC, University of Amsterdam & University of Liverpool.

R. Parikh, 1985, 'The Logic of Games and its Applications', *Annals of Discrete Mathematics* 24 (1985), 111 – 140.

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

Ph. Rohde, 2005, *On Games and Logics over Dynamically Changing Structures*, Dissertation, Institute of Informatics, RWTH Aachen.

R.A.M. van Rooij, 2004, 'Signalling Games Select Horn Strategies', *Linguistics and Philosophy* 27, 493 – 527.

H. Rott, 2007, 'Information Structures in Belief Revision'. To appear in P. Adriaans & J. van Benthem, eds., *Handbook of the Philosophy of Information*, Elsevier Science Publishers, Amsterdam.

O. Roy, 2007, *Logic, Intentionality, and Decision*, Dissertation, Institute for Logic, Language and Computation, University of Amsterdam.

K. Segerberg, 1995, 'Belief Revision from the Point of View of Dynamic Doxastic Logic', *Bulletin of the IGPL* 3, 534 – 553.

F. Veltman, 1985, *Logics for Conditionals*, Dissertation, Philosophical Institute, University of Amsterdam.

O. Weinberger, 1965, *Der Relativisierungsgrundsatz und der Reduktionsgrundsatz – zwei Prinzipien des dialektischen Denkens*, Nakladatelství Ceskoslovenské akademie Ved, Prague.