Chapter 1 INFORMATION DYNAMICS, RATIONAL AGENCY, AND INTELLIGENT INTERACTION

1.1 Logical Dynamics

Human life is a history of millions of actions flowing along with a stream of information. We plan our trip to the hardware store, decide on marriage, rationalize our foolish behaviour last night, or prove an occasional theorem, all on the basis of what we know or believe. Moreover, all this activity takes place in constant interaction with others, and it has been claimed that what makes humans so unique in the animal kingdom is not our physical strength, nor our isolated powers of deduction, but rather our planning skills in social interaction, with the Mammoth hunt as an early example, and legal and political debate as a manifestation. It is this intricate cognitive world which I take to be the proper domain of logic, as the study of the invariants underlying these informational processes. In particular, my program of *Logical Dynamics* (van Benthem 1991, 1996, 2001) calls for identification of a wide array of informational processes, and their explicit incorporation into logical theory, not as didactic background stories for the usual concepts and results, but as first-class citizens. One of the starting points in that program was a pervasive ambiguity in our language between *products* and *processes*. 'Dance" is an activity verb, but it also stands the product of the activity: a waltz or a mambo. "Argument" is a piece of a proof, but also an activity one can engage in, and so on. Logical systems as they stand are product-oriented, but Logical Dynamics says that both sides of the duality should be studied to get the complete picture. And this paradigm shift will send ripples all through our standard notions: for instance, 'natural language' will now be, not a static description language for reality, but as a dynamic *programming language* for changing cognitive states.

Two more recent alternative terms are informative here, each in its own way. 'Rational agency' stresses the transition from the paradigm of proof and computation performed by a single agent (or none at all) to the realities of many agents with abilities and preferences plotting their course through life. Incidentally, this same turn may be observed in computer science, which is no longer about lonely Turing Machines scribbling on their tapes, but about complex intelligent communicating systems which even have goals and purposes.

A second new term ‘intelligent interaction’ emphasizes the fact that cognitive powers are displayed at their best in many-mind, rather than single-mind settings – just as physics only really gets interesting when we no longer focus on single bodies searching for their ‘natural place’ in the Aristotelean sense, but on the Galilean-Newtonian view of a vast system of many bodies in the universe pulling at each other, form nearby and far away.

Now, this goes beyond the traditional agenda of the field of logic, as the study of ‘consequence relations’, or ‘formal systems’, or ‘provability and definability’. And I have to confront two groups of skeptics. Traditional logicians will feel this is going much too far, stretching ‘logic’ beyond breaking point – while losing what one colleague called the ‘nobility’ of abstract propositions and proofs, as completely disjoint from human strife and messiness. To this, I can only say that this whole book is clearly still logic in two senses. First, it uses existing systems in the field, but it shows that their reach is in fact much larger than what is usually realized, including revealing descriptions of social informational processes. Thus, dynamic logics can treat, say, *asking a question* to get information as a paradigmatic ‘logical’ action, just as much as the standard act of performing an inference. But also, the style of research is clearly with the mind set of a ‘logician’: to see that, just read this book. In fact, I would also be happy to define ‘logic’ in the same dynamic social style as a Dutch colleague of mine once did when challenged about his newfangled interests in formal semantics. Was this still linguistics, or had he left the flock? Here is what he said: “’Linguistics’ is that set of topics which is studied by active linguists. I am an active linguist. Therefore, what I am doing is linguistics.” A field is the set of themes pursued by its lively practitioners. That process-oriented answer, though of course less serious than my earlier one, is at least consistent with the agenda I just sketched.

The other group of skeptics are practitioners of successful other disciplines. The area of rational agency and intelligent interaction may be the logician’s Promised Land, but it is hardly virgin territory. Like Palestine in the Old Testament, it is already densely settled by other nations, such as philosophers, social scientists, or economists, worshipping other gods, such as probability or game theory. Can we just move in and take their lands? My answer is of course negative: we cannot, and we should not. My points are just these. Rational agency and intelligent interaction are complex and notoriously difficult subjects.

To grasp them fully, we need all the help we can get. In particular, I think that logic has something fresh to offer here, in terms of notions and results, in addition to the insights already in place. But personally, I see no favoured position for the logical approach I am pursuing here. Monotheism is just a bad idea, also in science. A philosophical essay on rational agency, or a game-theoretic probabilistic analysis of a common social scenario, may be just as insightful, and we should use logic only where it adds value. But what I do expect to see are fruitful marriages between logical and other points of view – and current contacts between logic and probability, and logic and game theory are promising examples.

So much for grand aims. The following simple examples will illustrate what we are after, and each adds a theme to our view of rational agency. We then summarize the resulting research program, followed by a brief description of the actual contents of this book.

1.2 The restaurant: logic in the eyes of a child

The Amsterdam Science Museum *NEMO (*http://www.nemo-amsterdam.nl/) organizes regular 'Kids' Lectures on Science'. Imagine 60 children aged around 8 in a small amphitheatre – with parents present in the wings, but not allowed to speak. In February 2006, it was my turn – and while preparing, I got more and more worried. How does one talk logic to such an audience? Was there *anything* in common between children that age and the abstractions driving one's university career? How to even start? My first question was this:

> In a restaurant, your Father has ordered Fish, your Mother ordered Vegetarian, and you have Meat. Out of the kitchen comes some new person with the three plates. What will happen?

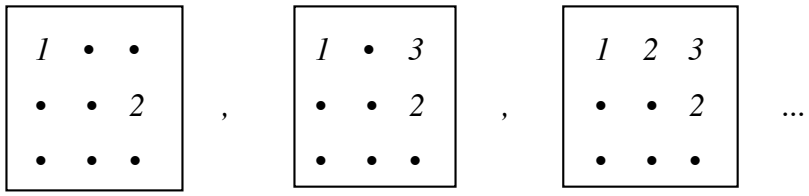
The children got excited, many little hands were raised, and one said: "He asks who has the Meat". "Sure enough", I said: "He asks, hears the answer, and puts the plate on the table. Now what happens next?" Children said "He asks who has the Fish!" Then I asked once more what happens next? And now one could see the Light of Reason suddenly start shining in those little eyes. One girl shouted: "He does not ask!" Now, *that* is logic … After that, we played a long string of scenarios, including card games, Master Mind, *Sudoku*, and even card magic, and we discussed what best questions to ask and conclusions to draw.

Through the eyes of the children, one also gets a fresh 'innocent' look at one's own field.

In my view, the Restaurant is about the simplest realistic logical scenario. And here is what we see. When the waiter puts that third glass without asking, [1] you see a logical inference in action 'in broad daylight': the information in the two answers received allows the waiter to deduce where the third one must go, via a valid propositional schema

$$A \vee B \vee C, \neg A, \neg B \Rightarrow C.$$

This sounds familiar, and one can now go on to the usual explanation of what logic is about, citing many further scenario's where this sort of valid inference takes place. For a concrete illustration of similar patterns of reasoning, think of successive stages in the solution of a 3x3 Sudoku puzzle, produced by applying the two basic rules that each of the 9 positions must have a digit, but no digit occurs twice on a row or column:



Each successive diagram displays a bit more information about the solution, which is already determined by the initial placement of the digits *1, 2*. Thus, information is brought to light in logical inference in a process of what may be called deductive *elucidation.*

So far, we have the usual door to logic here – but let us try to look with fresh eyes.

1.3 Entanglement of logical core tasks: inference and observation

The restaurant example cries out for a new twist. There is a natural unity to the scenario. The waiter first obtains the right 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, both informational processes are on a par, and both should be within the compass of logic, which is about information flow in general,

[1] When hearing this example, a former President of Amsterdam University gave me a warning: "Johan, you should be more careful and avoid low-class establishments. When *I* order something, I am not paying all that money to just have my glass put down, while the others get a question."

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?

How to do it? To model the information changes at our table, a helpful metaphor is some abstract form of *computation*. During a conversation, information states of people – singly, and in groups – change over time, in a systematic way triggered by communicative events. So we need a universe of states and possible transitions between them. Consider the information flow in the restaurant 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 plates 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:

This is the proto-typical process of semantic updates of the current *information range*, where new information is produced by making observations that rule out possibilities.

Note that the first two updates have already zoomed in on the actual situation. Nothing would change if we now tried to update with the conclusion, and this explains why no third question is needed. We have reached the true world, and now one just needs to spell it out. This is where inference kicks in, elucidating what the world looks like through another set of dynamic changes, now on syntactic descriptions of the world, as we have just seen.

Actually, this description is a bit disingenuous: we are already at a fault zone in logic here. Our scenario illustrates the folklore sense in which valid conclusions 'add no information' to what is already given by the premises. But then, what is the *point* of drawing a logical inference at all? To explain this, it is often said that inferences 'unpack' the information which the semantic update gave us only implicitly. But then, the latter is about a different notion of 'explicit information' on top of the semantic one. Now, while there are elegant logics for observation-based update of semantic information, there just is no generally accepted account of what information is produced by inference (cf. van Benthem &

Martinez 2008). Chapter 4 of this book has a proposal, working in analogy with semantic update – but still, this aspect of the entanglement is already a serious issue in its own right. Thus, our seemingly trivial scenario raises deep and difficult questions. And that is good.

Putting these things together, the *dynamics* of various kinds of informational actions becomes an obvious target for logical theory. Of course, to make this work, we must also give an account of the underlying *statics*: the information states that the actions work over. Both are precisely what will happen in the various chapters of this book. And as a first step toward this program, we have identified the first level of skills that rational agents have:

> *their powers of* inference*, and their powers of* observation*, resulting in information updates which change what they currently know*.

1.4 Information about others and public social dynamics

Here is another striking feature to the information flowing in the restaurant. Questions and answers typically involve more than one agent, and hence the dynamics is *social*, having to do also with what people (come to) know about each other. This higher-order knowledge about others is crucial to human communication and interaction in general.

Questions and answers Take just one simple question/answer example, the ubiquitous building block 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 much more is 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. [2] Moreover, your 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 iterations like this, they achieve so-called *common knowledge* of the relevant facts in the

[2] Naturally, all such presuppositions are off in a classroom with a teacher questioning students.

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. [3] 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.

So, logical activity is interactive, and its theory should reflect this. Some colleagues find this alarming, as social aspects are reminiscent of gossip, status, and Sartre's "Hell is the Others". The best way of dispelling such fears may be a concrete example. Here is one, using a card game, a useful 'normal form' for studying information flow in logical terms.

The Cards (van Ditmarsch 2000) Three cards 'red', 'white', 'blue' are distributed over three players: *1, 2, 3*, who get one each. Each player sees her own card, but not the others. The real distribution over *1, 2, 3* is *red, white, blue*. Now a conversation takes place (this actually happened during the *NEMO* children session, on stage with thee volunteers):

2 asks 1	*"Do you have the blue card?"*
1 answers truthfully	*"No"*.

Who knows what then, assuming the question is sincere? Here is the effect in words:

> Assuming the question is sincere, *2* indicates that she does not know the answer, and so she cannot have the blue card. This tells *1* at once what the deal was. But *3* did not learn, since he already knew that *2* does not have blue. When *1* says she does not have blue, this now tells *2* the deal. *3* still does not know even then; but since she can perform the reasoning just given, she knows that the others know.

[3] If I find your pin code and bank account number, I may empty your account, if I know that you do not know that I know all this. If I know you know that I know, I will not. Successful criminal activity is triggered by fine iterated epistemic distinctions: that is why it is so hard.

We humans go successfully through this sort of reasoning in many settings, allowing for different knowledge for different agents. In Chapters 2, 3 we will analyze this information flow in detail, using concrete semantic diagrams to keep our intuitions straight.

These scenarios can be much more complex. Real games of 'who is the first to know' arise by restricting possible moves for players, and we will consider such scenarios later on. Also, public announcements raise the issue of the reliability of the speaker, and as a simple case, the famous logic puzzles of meetings with Liars and Truth-Tellers. Our systems will also be able to deal with these in a systematic manner, though 'asking the right question' to separate one agent type from another is already a subtle manner of design. Logic of communication is not easy, but our point here is that it is about well-defined issues.

Thus, we have a second major aspect of rational agents in place as a challenge to logic:

> *their* social powers *of mutual knowledge and communication.*

1.5 Partial observation and differential information

We now have the beginnings of a much broader set of questions for logical analysis. Clearly, public announcement is just one way of creating new information. The reality in the above games, and indeed most social situations, is that information flows differentially, with partial observation by agents. When I draw a card from the stack, I see which card I am getting, you do not, though you may know it is one of a certain set: you get *some* information, but I get more. When you take a peep at my card, you learn something while cheating me, degrading my knowledge of the current state of the game into mere belief. When you whisper something in your neighbour's ear during my talk, this is a form of public announcement in a subgroup, where I and the others do not catch what you are saying, and I may not even notice that any information is being passed at all.

Modeling this kind of information flow in a systematic manner is much more complicated than public announcement, and goes beyond existing logical systems. The first satisfactory proposals were made only in the late 1990s, as we shall see in this book (cf. Gerbrandy 1999, Baltag, Moss & Solecki 1998, van Benthem, van Eijck & Kooi 2006) – and by now, we can model information flow in realistic parlour games like "Clue" (van Ditmarsch

2000), which have an intricate system of public and private moves. All this occurs in natural realities that we humans swim in every day, such as *electronic communication*:

> I send you an email, with the message *'P'*: a public announcement in the group {you, me},
> You reply with the message *'Q'* with a ***cc*** to others: public announcement to larger group.
> I respond with *'R'* with a 'reply-to-all' plus a ***bcc*** to some further agents:

In the third round, we have one of these partly hidden acts again: using *bcc* I have made a public announcement to different groups, but one group does not know that others were included. I challenge the reader to model the information flow from start to finish in this quite common episode: it is not easy, and a large pat of the thesis Ji 2004 was devoted to just this issue. Indeed, after a few rounds of consecutive *bcc* messages to different groups, it becomes very hard to keep track of who is supposed to know what. And that makes sense: differential information flow is complex, and so is understanding social life. Chapter 5 of this book will develop the theory behaind this in much greater detail.

Indeed, there are many thresholds one wants to understand here. Using *bcc* is not necessarily misleading to agents who know that it is a possible occurrence in the system. This is like games where the official rules printed on the cardboard box allow for moves that favour some players over other informationally. A border line is crossed when we allow lying and cheating beyond the expected. But even this seems a crucial skill in our lives: just imagine the violence and bloodshed that would result in academic life if we told our colleagues honestly what we think of the quality of their work. In this connection, parents think that their children are innocent little angels because they speak the truth all the time. But really, that is just a sign of lack of processing power and immaturity: rational agents at full capacity can handle mixtures of lies and truths with elegance and ease.

Thus, we have a further twist to our account of rational agents:

> different observational access *and processing differential information flow*.

This may seem terribly 'engineering-like'. Are not we supposed to study Knowledge and Truth and heroic lonesome agents achieving this grand harmony between mind and the facts? Who cares about the sordid realities of chatting, lying, and social manoevering? Well, first, differential information is a great good: we do not tell everyone everything, and

this is crucial to keeping things civilized, and efficient. Moreover, given the fact that all successful human activity is social, from hunting cave bears to mathematical research, it is amazing how understanding the fascinating social information flow which keeps all this functioning has been such a low priority of logicians and philosophers for such a long time.

1.6 Epistemic shocks: self-correction and belief revision

So far, we considered information flow and knowledge. Now, 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. But 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 occur with 'warm-blooded agents', who are opinionated, make mistakes, but who subsequently *correct* themselves. [4] Thus, rationality is about the dynamics of *being wrong* just as much as about that of being right, through belief revision, or *learning* by giving up old beliefs.

In a very concrete setting, revision comes to the fore in conversation, one of our key examples so far. 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.

Here, events become more delicate than what we had so far with recording information flow through observation. For instance, our knowledge can never be shaken by receiving true information, but our beliefs can, when we learn new facts contradicting what we considered most plausible so far. Feeling that an earthquake is hitting the Stanford campus, I no longer believe that a short bike ride through the night will get me home in 10 minutes. But there is much more for logic to keep straight in this area. For instance, the following nasty scenario has been discussed by computer scientists, philosophers, and economists in the 1990s. Even true beliefs can be sabotaged through true information:

---

[4] 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…

Misleading with the Truth You know you finished *3d, 2d,* or *1st* in the election, and you think lower outcomes more plausible. You also know that being *2d* will make your chances of high office small ('dangerous heavy-weight'). In fact you were *1st*. I know this, but only tell that you are not *3d*: and you become quite unhappy. Why?

Initially, you believe that you may get high office by way of compensation, because being *3d* is the most plausible outcome. So, you have a true belief, though for the wrong reason. Now you learn the true fact that you are not *3d*, with being *2d* becoming the most plausible world. But then, you have now come to believe, falsely, that you will not have any success – something you would not believe if you knew that you won the election. Our logics of belief revision in Chapter 6 of this book can easily deal with such scenarios, as well as with others with a 'softer touch', where the incoming information merely makes certain worlds less or more plausible, without ever removing any world entirely from consideration.

Thus, in addition to monotonic knowledge update, we have identified another, more 'jumpy', but equally important feature of rational agents:

their capacity for hypothesizing, being wrong, and then correcting themselves.

In many settings, these capacities seems the more crucial and admirable human ability. A perfectly healthy body is great, but lifeless, and the key to our biological performance is our immune system responding to cuts, bruises, and diseases. Likewise, I would say that flexibility in beliefs is essential: and logic is all about the immune system of the mind!

1.7 Planning for the longer term

So far, we have discussed single moves that rational agents make in response to incoming information, whether knowledge update or belief revision. But in reality, these single steps make sense only as part of longer processes through time. A conversation is a sequence of steps, each responding to earlier ones, and usually directed toward some goal, and the same is true for many games, and social activities in general. There is relevant structure at this level, too, and as usual, it is high-lighted by well-known puzzles. Here is an evergreen:

The Muddy Children (Fagin et al. 1995): After having played outside, two of three children have got mud on their foreheads. They can only see the others, so they do

> not know their own status. [5]Now their Father comes along and says: "At least one of you is dirty". He then asks: "Does anyone know if he is dirty?" Children answer truthfully, round by round. As questions and answers repeat, what happens?

One might think that nothing happens, since the father just tells the children something they already know – the way parents tend to do –, viz. that there is at least one dirty child. But in reality, he does achieve something significant, making by turning this fact into common knowledge. Compare the difference between every colleague knowing that your partner is unfaithful: no doubt unpleasant, but maybe still manageable, with this fact being common knowledge, including everyone knowing that the others know, etcetera: the shame at department meetings becomes unbearable. Keeping this in mind, here is what happens:

> Nobody knows in the first round. But in the next round, each muddy child can reason like this: "If I were clean, the one dirty child I see would have seen only clean children, and so she would have known that she was dirty at once.
> But she did not. So I must be dirty, too!"

Note that this scenario is about what happens in the long run: with more children, common knowledge of the muddy children arises after more rounds of ignorance announcement, and in fact, in the next step, the clean children will know that they are clean.

Notice also that there is obvious formal structure to this scenario. The nstruction to the children looks like a little computer program:

> REPEAT (IF don't know your status THEN say you don't know ELSE say you know).

And this is no coincidence. Conversation involves *plans*, and these plans have the same 'control structure' for actions found in computer programs: choice, sequential composition, and iteration of actions. On top of this, the Muddy Children even display a sophisticated *parallel composition* of actions, since they answer simultaneously. Thus, we see that actions may be composed and structured to achieve long-term effects – and this, too will be an aspect of our logics. But for the moment, we note this:

---

[5] This observational access is the inverse of our earlier card games, but formally very similar.

Rational agency involves planning in longer-term scenarios*, and its quality also lies in the ways that agents compose their individual actions in larger wholes.*

We will return to this longer-term perspective on agency and interaction, but let us first identify another aspect which is crucial to understanding the 'driving force' behind it.

1.8 Preference and goals

Just answering 'a simple question' is rare. Behind every question posed to us, there lies a *why*-question: what does this person want, and in what sort of conversation, or course of action am I entering? Pure informational orientation is rare, and our informational activities tend to live in an ether of preferences and evaluation of actions and their outcomes. This is not just because of monetary gain or emotional response. 'Making sense' of an interaction does not just involve meaning and information, but also mutually getting clear on the goals of everyone involved. This brings in another level of agent structure: crucially,

logic of rational agency involves preferences *between situations and actions.*

Our preferences determine what actions we take, and knowing your preferences allows me to make predictions about what is going to happen. And if you think this never happens in pure fields like mathematics, just observe how a referee will judge your paper on its 'interest' rather than just plain truth, where 'interesting' depends on the preferences of a scientific community. Humans just cannot separate information from evaluation, and this probably reflects some deep entanglement of the cognitive and emotional system inside our brains, for good evolutionary reasons. But even leaving that more sweeping perspective aside, the notions of preference and utility have long been acknowledged as fundamental.

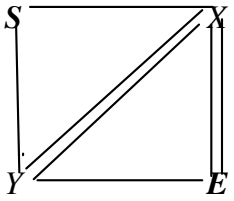
Now, while preference has been studied extensively in decision theory and game theory, it has been much more marginal in logic. There is indeed a field called 'preference logic', which has led a somewhat separate life. In this book, we will incorporate it, and show how it fits very well with logical analysis of information update and belief revision.

1.9 Games and intelligent interaction

Temporal perspective also comes in with the next crucial feature of rational agents, their responding to, and influencing others. Even the simplest form of conversation, our original

example, involves choosing assertions depending on what others say. This interactive aspect means that dynamic logics must eventually come to turn with *games*.

True interaction and games To sample the spirit of interaction, consider the following 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 the 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 the escape *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 further cutting
> be guided in a straightforward manner 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. [6]

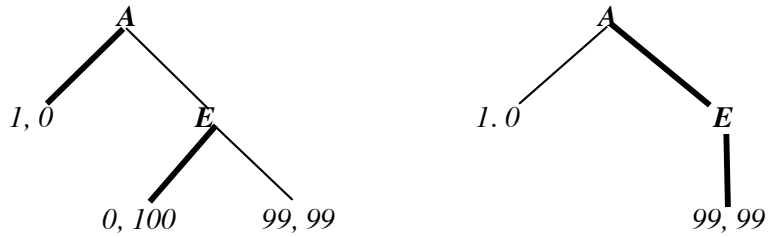
Models of learning This two-agent example is also meant to show how hard it is sometimes to abandon more classical single-agent views. Formal Learning Theory concentrates on single-agent settings where a student forms hypotheses on the basis of some input stream of evidence: there is a Student, but no Teacher, unless we think of 'Nature' as a passive teacher doing the minimum of uniform presentation without

[6] Rohde 2005 shows that solving 'sabotage graph games' like this is *Pspace*-complete. The reader will get an even better feel for the complexity of interaction by considering the following variant of our teaching game. 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, in the same graph?

adjustment to reach a comfortable retirement. But the realities of teaching and learning are social, with Students and Teachers responding to each other, and learning is largely a social process, where we even learn at two levels: concerning the outcomes of the game: the 'knowledge' imparted, but also about successful strategies, the 'know-how' or the 'skills'.

Logic and game theory With multi-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*: that is, one of the two players has a winning strategy. This result applies immediately to our teaching game, and it explains why in each finite graph of this sort, either Student or Teacher has a winning strategy (for details, cf. Chapter 9 below).

But real game theory only arises when we add players' preferences and evaluation of outcomes as before. In that setting, the solution procedure extending Zermelo's is called *Backward Induction*. 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'. Here is an example:



(a) (b)

The thick black lines in the game tree (a) indicate what 'rational' players should do, describing the only 'subgame-perfect' Nash equilibrum of this game. Now, there are some obvious questions here, since this is obviously a 'bad equilibrium' representing a socially undesirable outcome, since *(1, 0)* makes both players worse off than the outcome *(99, 99)*. Thus, we need to carefully spell out, and reassess, all assumptions behind the usual solution procedures for games, even those which have already gained their inventors Nobel Prizes. And here is where dynamic logics of agency can help, as we will see in Chapters 8, 9.

Moreover, our logics are not just about preserving the status quo. They may also suggest quite different takes on these scenarios. In Chapters 9, 14, we will consider a logic of

promises that change a current game through public announcements of intentions. ***E*** might promise that she will not go left, changing game (a) to game (b) – and the new equilibrium *(99, 99)* results, making both players better off. Van Benthem 2008 proposes alternatives to Backward Induction in terms of history-oriented versions of games, where players remind themselves of the *legitimate rights of others*, because of 'past favours received'.

Finally, games are not just some analytical tool. They are a ubiquitous human activity, which is stable across cultures, and seems to serve a lot of needs, ranging from gentle elegant waste of time to training for crucial skills. Our conclusion is that

> *a full logical understanding of rational agency and intelligent interaction*
> *requires a logical study of* games*, as a crucial model for human behaviour.*

This theme will be considered more extensively in the companion volume van Benthem, to appear, but in Chapters 9, 13, 14, we take it up in some detail, including further mixtures with our earlier themes, such as 'knowledge games' about being the first to know the actual world, or logical structures in games with partial observation and imperfect information.

1.10 The long run, the big bulk, and the large numbers

It is time to also set some limits to what we are trying to achieve here.

The 'Grand Stage' of temporal logic Our study of rational agency is about individual agents engaged in meaningful tasks which they can control from beginning to end. But at some aggregation level of time or space, 'large-scale phenomena' will take over. This is the case over time, when we study a potential infinity of events, and then, our dynamic logics of agency will run into temporal logics that describe a sort of Grand Stage of branching histories, where all individual scenarios take place. Such logics have been proposed in various areas of philosophy and computer science, and one can trace their origins even in the world literature, witness Borges' famous story 'The Garden of Forking Paths'. We will discuss this junction with other branches of logic and process theory in Chapter 11.

Groups, social structure, and collective agency Likewise, single agents need not just interact on their own: typically, they also form groups and other collective agents, whose behaviour may not be totally reducible to that of individual members. For instance, social

choice theory is about groups deciding their preferences on the basis of individual preferences of their members. Or, players in games can form coalitions membership of which may influence their behaviour. We saw some of this in the notion of common knowledge, which is about the degree of being informed inside a group. But there are more junctions to be made, and in Chapter 10 we will show how our dynamic logics interface with group behaviour, and even may help provide an underpinning to social choice theory, in the form of a 'micro-theory' of information-based rational deliberation.

Probability and statistics Finally, on top of the logical micro-structure that we will focus on, macro-structure has its own notions appropriate to its laws, just as physics has pressure and temperature at a phenomenological level above the mechanics of colliding particles. Probabilistic methods then come to the fore, and they are indeed prominent in game theory with the use of 'random strategies' to describe optimal play, or the study of opinion flow through large communities, which can be studied as a physical transport phenomenon. We will discuss some connections between our dynamic logics and probabilistic update in Chapter 7, though this book will not really explore this interface at any great depth. Even so, we conclude with one example of probabilistic considerations, just to round off our presentation – and show where other hands may have to take over.

Games with probability The following story is from a column by Marilyn Vos Savant, San Francisco Chronicle, March 2002. "A stranger walks up to you in the municipal library (where you are reading this book) and offers to play the following game:

> You both show heads or tails. If you both show heads, she pays you $1,
> if both show tails, then she pays $3, while you must pay her $2 in
> case you show different things. Is this game fair?"

Is not your expected value $1/4 \bullet (+1) + 1/4 \bullet (+3) + 1/2 \bullet (-2) = 0$? In her response the next week, Vos Savant pointed out the game was unfair to you with *repeated play*. The stranger can then play heads two-thirds of the time, which would give you an average pay-off of

$$2/3 \bullet (1/2 \bullet (+1) + 1/2 \bullet (-2)) + 1/3 \bullet (1/2 \bullet (+3) + 1/2 \bullet (-2)) = -1/6.$$

But what if I had played a different counter-strategy against this behaviour of the stranger, viz. 'Heads all the time'? Then my expected value would have been *2/3•(+1) + 1/3•(-2) = 0*. So, what *is* the fair value of this game – and should I engage in it?

This requires reasoning with probabilistic strategies, or bulk behaviour over time, with players i playing actions σ_i with probability p_i, where the values p_i for all players add up to 1. Von Neumann and Nash showed that all finite strategic games have equilibria in such mixed strategies, where no player can improve her expected value by unilaterally deviating. Computing the solution for the Library Game in this way, we find that you should show Heads with probability *5/8,* and so should the stranger. The expected value of this game for you is *–1/8*, making it unfavourable. Thus, probability theory adds subtlety to logic.

1.13 The program of 'Logical Dynamics' in a nutshell

Rational agency Collecting the preceding examples and considerations, we get a rich picture of rational agency and intelligent interaction as a field for logic. Meaningful logical tasks performed by human agents involve information flow with many entangled activities: inference, observation, questions, and communication. 'Logical Dynamics' says that all these activities should be first-class citizens, and logical theory should treat them on a par. This view goes beyond standard definitions of logic as being only about valid consequence. But this book does not abandon classical logical systems: to the contrary, it shows how they can describe more kinds of information flow than inference, when enriched with tools from dynamic logic. The resulting 'dynamic logics' will look less familiar, but they add subtlety and scope making it worth the effort of learning them. [7] Admittedly, all this means we are placing the logical study of *rational agency* at centre stage, instead of agent-free proof and computation. Moreover, we form new links to fields beyond the usual 'friends of logic: mathematics, philosophy, and linguistics. New interfaces are studies of agency and intelligent systems in computer science, game theory in economics, and the social sciences.

[7] Moreover, there is a side benefit to the 'Dynamic Turn'. We can often replace 'non-standard logics', which have sprung up in droves in recent years, by perfectly classical systems, once we identify the proper information-changing events, and make them an explicit part of the logic.

A historical pedigree Is all this merely new-fangled tinkering with the core values of logic? I do not think so. The ideas put forward here are ancient and obvious. 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, say, 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. [8] And the social interactive aspect of information flow is 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 patterns of confirmation and refutation between participants It has been claimed that logic arose out of political and legal debate in all its three main traditions: Chinese, Indian, and Western. [9] And this multi-agent interactive view has emerged anew in modern times. A beautiful case are the *dialogue games* of Lorenzen 1955, which explained logical validity in terms of a third pragmatic notion, different from both proof-theoretic and semantic intuitions, as the existence of a winning strategy for a proponent arguing the conclusion against an opponent granting the premises. Similar views may be found in Hintikka 1973, another pioneer of taking games and informational activities seriously inside logic. [10]

---

[8] At a 2005 Winter School, IIT Bombay (cf. Gupta, Parikh & van Benthem, eds., 2007), I made this point to students about the question if there was 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. Now I was down to asking experts, viz. the students. No answer was forthcoming, 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 needed a good deal of beer, we bought you three bottles."

[9] The Mohists in early China discuss the Law of Non-Contradiction as a principle of conversation: 'resolve contradictions with others', 'avoid contradicting yourself'. Cf. Zhang & Liu 2007.

[10] Again, we link logic and games in great detail in the companion book (van Benthem, to appear).

In our view, logic has been interactive all along, and in pursuing the Logical Dynamics of this book, we are merely scraping off some layers of convention. While thus reclaiming a broader traditional agenda, with formal systems a means but not the end, logic also becomes a central part of academic life, overflowing the usual disciplinary boundaries.

I.14 The chapters explained

In this book, we develop the general dynamic logic for rational agency outlined in the Introduction. This requires a number of steps, following the distinctions in activities and powers of agents that we have made. The first set of chapters is about *information structure and knowledge update*, where we take knowledge in the relaxed sense of what is true according to the agent's hard information. Our tools are standard epistemic logic (Chapter 2), its dynamified version called 'public announcement logic' (*PAL*; Chapter 3) which can deal with public communication or observation, and finally, its more sophisticated version called 'dynamic-epistemic logic' (*DEL*; Chapter 4). These systems are our paradigm for systematic analysis of definable changes in a current possible worlds model of the agents' information, with possible worlds taken in a relaxed sense, without metaphysical overtones. This methodology works much more broadly. A first illustration is the analysis in Chapter 5 for inference steps, procedural information, and actions turning implicit into explicit knowledge. Next, the approach is applied to *belief revision and self-correction* in Chapter 6, suitably extended to changes in plausibility orders, showing how revision can be dealt with without any ad-hoc theory. Chapter 7 is a digression, showing how similar ideas work for probabilistic update, with a more refined quantitative view of learning mechanisms. Next, we move to agents' goals and show how our techniques for qualitative plausibility change also apply to agents' *preferences* and ways of changing them, providing a unified account of information, belief, and goals (Chapter 8). Next, we emphasize two topics having to do with the social nature of the dynamics that we are after. Chapter 9 is about *multi-agent interaction*, adding 'program structure' to single update or revision steps to model plans and longer term scenarios – extending our dynamic logics with ideas from computer science and especially, game theory. Next, Chapter 10 considers *groups* as new logical agents, showing how the earlier systems lift to this setting, while studying new phenomena such as belief merge, as well as connections to social choice theory. Finally,

while Chapters 2 through 8 dealt with local steps of change in information, belief, or preference, and Chapter 9 looked at terminating scenarios like conversations or finite games, a longer-term temporal perspective is taken in Chapter 11, leading to embeddings of our dynamic logics in *temporal logics* of knowledge and belief in branching time, as well as merged topics such as logics of protocols. The remaining Chapters 12, 13, 14, 15 show how the logical dynamics developed here applies to a range of issues in philosophy, computer science, game theory, and cognitive science – if only as a way of reinterpreting old questions in these fields, and adding fresh ones. Chapter 16 states our conclusions.