# 1 Logical dynamics, agency, and intelligent interaction

## 1.1 Logical dynamics of information-driven agency

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, 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 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 late manifestation. It is this intricate cognitive world that I take to be the domain of logic, as the study of the invariants underlying these informational processes. In particular, my programme 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 programme was a pervasive ambiguity in our language between *products* and *activities* or processes. 'Dance' is an activity verb, but it also stands for 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 a dynamic programming language for changing cognitive states.

Recent trends have enriched the thrust of this action-oriented programme. 'Rational agency' stresses the transition from the paradigm of proof

and computation performed by a single agent (or none at all) to agents with abilities, goals, and preferences plotting a meaningful course through life. This turn is also clear in computer science, which is no longer about lonely Turing Machines scribbling on tapes, but about complex intelligent communicating systems with goals and purposes. Another recent term, 'intelligent interaction', emphasizes what is perhaps the most striking feature here, the role of *others*. Cognitive powers show at their best in many-mind, rather than single-mind settings – just as physics only gets interesting, not with single bodies searching for their Aristotelian natural place, but on the Newtonian view of many bodies influencing each other, from nearby and far.

## 1.2 The research programme in a nutshell

What phenomena should logic study in order to carry out this ambitious programme? I will first describe these tasks in general terms, and then go over them more leisurely with a sequence of examples. A useful point of entry here is the notion of *rationality*. Indeed, the classical view of humans as 'rational animals' seems to refer to our reasoning powers:

*To be rational is to* reason *intelligently*.

These powers are often construed narrowly as deductive skills, making mathematical proof the paradigm of rationality. This book has no such bias. Our daily skills in the common sense world are just as admirable, and much richer than proof, including further varieties of reasoning such as justification, explanation, or planning. But even this variety is not yet what I am after. As our later examples will show, the essence of a rational agent is the ability to use information from many sources, of which reasoning is only one. Equally crucial information for our daily tasks comes from, in particular, observation and communication. I will elaborate this theme later, but right now, I cannot improve on the admirable brevity of the Mohist logicians in China around 500 BC (Zhang & Liu 2007):[1]

Zhi: Wen, Shuo, Qin 知 问 说 亲

*knowledge arises through questions, inference, and observation*.

[1] Somewhat anachronistically, I use modern simplified Chinese characters.

Thus, while I would subscribe to the above feature of rationality, its logic should be based on a study of all basic informational processes as well as their interplay.

But there is more to the notion of rationality as I understand it:

*To be rational is to* act *intelligently*.

We process information for a purpose, and that purpose is usually not contemplation, but action. And once we think of action for a purpose, another broad feature of rationality comes to light. We do not live in a bleak universe of *information*. Everything we do, say, or perceive is coloured by a second broad system of what may be called *evaluation*, determining our preferences, goals, decisions, and actions. While this is often considered alien to logic, and closer to emotion and fashion, I would rather embrace it. Rational agents deal intelligently with both information and evaluation, and logic should get this straight.

Finally, there is one more crucial aspect to rational agency, informational and evaluational, that goes back to the roots of logic in Antiquity:

*To be rational is to* interact *intelligently*.

Our powers unfold in communication, argumentation, or games: multi-agent activities over time. Thus, the rational quality of what we do resides also in how we interact with *others*: as rational as us, less, or more so. This, too, sets a broader task for logic, and we find links with new fields such as interactive epistemology, or agent studies in computer science.

I have now given rationality a very broad sense. If you object, I am happy to say instead that we are studying 'reasonable' agents, a term that includes all of the above. Still, there remains a sense in which mathematical deduction is crucial to the new research programme. We want to describe our broader agenda of phenomena with *logical systems*, following the methods that have proven so successful in the classical foundational phase of the discipline. Thus, at a meta-level, in terms of modelling methodology, throughout this book, the reader will encounter systems obeying the same technical standards as before. And meta-mathematical results are as relevant here as they have always been. That, to me, is in fact where the unity of the field of logic lies: not in a restricted agenda of 'consequence', or some particular minimal laws to hang on to, but in its methodology and modus operandi.

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

## 1.3 Entanglement of logical tasks: inference, update, and information flow

The Amsterdam Science Museum *NEMO* (www.nemo-amsterdam.nl/) organizes regular 'Kids' Lectures on Science', for some sixty children aged around eight in a small amphitheatre. In February 2006, it was my turn to speak – and my first question was this:

> *The Restaurant* 'In a restaurant, your Father has ordered Fish, your Mother ordered Vegetarian, and you have Meat. Out of the kitchen comes some new person carrying 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. What happens next?' The 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.

***Two logical tasks*** The Restaurant is about the simplest scenario of real information flow. And when the waiter places that third plate without asking, you see a logical inference in action. The information in the two answers allows the waiter to infer (implicitly, in a flash of the 'mind's eye') where the third plate must go. This can be expressed as a logical form

$$A \text{ or } B \text{ or } C,\ \text{not-}A,\ \text{not-}B \Rightarrow C$$

One can then tell the usual story about the power of valid inference in other settings. With this moral, the example goes back to Greek Antiquity. But the scenario is much richer. Let us look more closely: perhaps, appropriately, with the eyes of a child.

To me, the Restaurant cries out for a new look. There is a natural unity to the scenario. The waiter first obtains the right information by asking questions and understanding answers, acts of *communication* and perhaps *observation*, and once enough data have accumulated, he *infers* an explicit solution. Now on the traditional line, only the latter step of deductive elucidation is logic proper, 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. Asking a question and grasping an answer is just as logical as drawing an inference. And accordingly, logical systems should account for both of these, and perhaps others, as observation, communication, and inference occur entangled in most meaningful activities.

***Information and computation*** And logic is up to this job, if we model the relevant actions appropriately. Here is how. To record the information changes in the Restaurant, a helpful metaphor is *computation*. During a conversation, information states of people – alone, and in groups – change over time, in a systematic way triggered by information-producing events. So we need a set of information states and transitions between them. And as soon as we do this, we will find some fundamental issues, even in the simplest scenarios.

***Update of semantic information*** Consider the information flow in the Restaurant. The intuitive information states are sets of 'live options' at any stage, starting from the initial six ways of giving three plates to three people. There were two successive *update actions* on these states, triggered by the answers to the waiter's two questions. The first reduced the uncertainty from six to two options, and the second reduced it to *1*, i.e., just the actual situation. Here is a 'video sequence' of how the answers for Meat and Fish would work in case the original order was ***FMV*** (fish for the first person, meat for the second, vegetarian for the third):

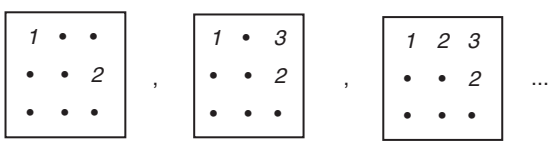

This is the common sense process of semantic update for the current *information range*, where new information is produced by events that rule out possibilities. In Chapters 2 and later, we will call this elimination scenario a case of 'hard information', and typical events producing it are public announcements in communication, or public observations.

***Inference and syntactic information*** The first two updates have zoomed in on the actual situation. This explains why no third question is needed. But then we have a problem. What is the *point* of drawing a logical conclusion if it adds no further information? Here, the common explanation is that inferences 'unpack' information that we may have only implicitly. We have reached the true world, and now we want to spell it out in a useful sort of code. This is where inference kicks in, elucidating by means of linguistic description what the world looks like:

(6) ***update 1 >*** (2) ***update 2 >*** (1) *inference to 'full arrangement'*

This sounds fine, but it makes sense only when we distinguish two different notions of information: one 'explicit', the other 'implicit' (van Benthem & Martinez 2008). Now, while there are elegant logics for semantic update of the latter, there is no consensus on how to model the explicit information produced by inference. Formats include syntactic accumulation of formulas, but more graphical ones also make sense. For instance, here is how propositional inferences drive stages in the solution of puzzles:

*Example* Take a simple 3×3 Sudoku diagram, produced by applying the two rules that 'each of the nine positions must have a digit', but 'no digit occurs twice on a row or column':




Each successive diagram displays a bit more about the unique solution (one world) determined by the initial placement of the digits *1, 2*. Thus, explicit information is brought to light in logical inference in a process of what may be called deductive *elucidation*. Chapter 5 of this book will make a more systematic syntactic proposal for representing the dynamics of inference, that works in tandem with semantic update. For now, we just note that what happened at the Restaurant already involves a basic issue in the philosophy of logic (cf. Chapter 13): capturing and integrating different notions of information.

Putting things together, the *dynamics* of various kinds of informational actions becomes a target for logical theory. But to make this work, we must,

and will, also give an account of the underlying *statics*: the information states that the actions work over. As a first step toward this programme, 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 that change what they currently know.*

## 1.4 Information about others and public social dynamics

Another striking feature to the information flow in the Restaurant are the questions. Questions and answers typically involve more than one agent, and their 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 'Yes/No' question followed by a correct answer, a ubiquitous building block of interaction. Consider the following dialogue:

> *Me*: 'Is this Beihai Park?'
> *You*: 'Yes.'

This conveys facts about the current location. But much more is going on. By asking the question 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 you may know the answer, again under normal circumstances.[2] Moreover, your answer does not just transfer bare facts to me. It also achieves that you know that I know, I know that you know that I know, and in the limit of such iterations, it achieves *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.[3] Keeping track of higher-order

[2] All such presuppositions are off in a classroom with a teacher questioning students. The logics that we will develop in this book can deal with a wide variety of such scenarios.

[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. But if I know that you know that I know,

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').[4] Indeed, the ability to move through an informational space keeping track of what other participants know and do not know, including the crucial ability to switch and view things from other people's perspectives, 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. It is like the Restaurant in some ways, but with a further layer of higher-order knowledge.

***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 three 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 stated 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 the deal; but since he can perform the reasoning just given, he does know that the others know it.

We humans go through this sort of reasoning in many settings, with different knowledge for different agents. In Chapters 2, 3, we will analyse this information flow in detail.

I will not touch your account. Crime is triggered by fine iterated epistemic distinctions: that is why it is usually better left to experts.

[4] Cf. Hendricks (2005), Verbrugge (2009), van Rooij (2005), and many other sources.

These scenarios can be much more complex. Real games of 'who is the first to know' arise by restricting possible questions and answers, and we will consider game logics later on. Also, announcements raise the issue of the reliability of the speaker, as in logic puzzles with meetings of Liars and Truth-Tellers. Our systems will also be able to deal with these in a systematic way, though separating one agent type from another is often a subtle manner of design. Logic of communication is not easy, but 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*.

Actually, these powers involve more than pure information flow. Questions clearly have other uses than just conveying information: they define *issues* that give a purpose to a conversation or scientific investigation. This dynamics, too, can be studied per se, and Chapter 6 will show how to deal with 'issue management' within our general framework.

## 1.5 Partial observation and differential information

The social setting suggests a much broader agenda for logical analysis. Clearly, public announcement as we saw in the Restaurant or with the Cards is just one way of creating new information. The reality in many games, and 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: getting *some* information. When you take a peek at my card, you learn something by cheating, degrading my knowledge of the current state of the game into mere belief. When you whisper in your neighbour's ear during my talk, this is a public announcement in a subgroup – where I and others need not catch what you are saying, and I may not even notice that any information is being passed at all.

Modelling such information flow 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 Chapter 4. By now, we can model information flow in parlour games like

*Clue*, that have an intricate system of public and private moves. All this occurs in natural realities all around us, such as *electronic communication*:

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

In the third round, we have a partly hidden act again: my *bcc* made an announcement to some agents, while others do not know that these were included. The information flow in this quite common episode is not simple. After a few rounds of *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.

There are intriguing thresholds here. Using *bcc* is not misleading to agents who know that it is a possible event in the system. A further step is *cheating*. But even judicious lies seem a crucial skill in civilized life. Our angelic children are not yet capable of that, but 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 the powers of rational agents:

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

This may seem mere engineering. Who cares about the sordid realities of cheating, lying, and social manoeuvring? Well, differential information is a great good: we do not tell everyone everything, and this keeps things civilized and efficient. Indeed, most successful human activity is social, from hunting cave bears to mathematics. And a crucial feature of social life is organization, including new procedures for information flow. Even some philosophy departments now do exams on Skype, calling for new secret voting procedures on a public channel. What is truly amazing is how this fascinating informational reality has been such a low priority of mainstream logicians and epistemologists for so long.

## 1.6 Epistemic shocks: self-correction and belief revision

So far, we considered information flow and knowledge. It is time for a next step. Agents who correctly record information from their observations, and industriously draw correct conclusions from their evidence, may be rational