Intelligent Interaction: dynamic trends in today's logic

Johan van Benthem, Amsterdam & Stanford

The original version of this text appeared as 'L'Art et la Logique de la Conversation', "Dossier Logique", Éditions *Pour la Science*, 2005, Paris, 68 – 73. The current version has been updated to include some recent developments. We show how modern logic attempts to deal with a wide range of intelligent interaction, crossing between academic disciplines from the humanities to the social and natural sciences. Topics include mathematical proof, information flow by communication or observation, and game-theoretic strategies, leading to a cognitive account of what makes us intelligent agents.

Logical proof steps When thinking of logic, most people have an image of inescapable inferences that force anyone, pauper or king, to accept their conclusions. Often these involve implications, of the form "if *A*, then *B*", or in logical notation, an arrow $A \rightarrow B$. A famous example is this rule, going back to Greek Antiquity:

Modus Ponens from given premises *A* and $A \rightarrow B$, draw the conclusion *B*

And once we see one such rule, we see others. A famous relative of Modus Ponens is

Modus Tollens from $A \rightarrow B$, and *not-B*, conclude to *not-A*.

The latter helps refute an opponent who claims *A*, by deriving some false implication *B* from *A*. Logical inferences like these involve steps on available evidence that are forced upon us. And though each single logical step may be obvious – and indeed, a bit boring – the cumulative force of many consecutive ones may acquire the force of a torrent. This calculus of assertions underlies simple everyday reasoning. Suppose you want to throw a party, respecting people's incompatibilities. You know that:

(a) John comes if Mary or Ann comes.
(b) Ann comes if Mary does not come.
(c) If Ann comes, John does not.

Can you invite people in a way that respects this? Logical mini-steps show the way:

By (c), if Ann comes, John does not. But by (a), if Ann comes, John does. This is a contradiction, so Ann does not come. But then, by (b), Mary comes. So, by (a) once more, John must come. Indeed, a party {John, Mary} satisfies all three requirements.

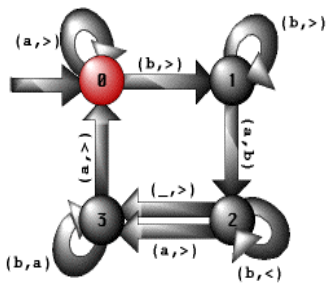
Indeed, millions of similar steps are used in modern automated theorem provers.

This proof-based view of logic, and its quest for absolute certainty, is intimately connected with the history of mathematics. Since Antiquity, a dominant paradigm of reasoning has been mathematical proof in axiomatic systems, as in Euclid's "Elements". This view of proof and axiomatic organization is also the backbone of the classical foundations of mathematics, where logicians have tried to show that the major mathematical theories are secure, i.e., free from provable contradictions.

Dialectical sources of logic But the origins of logic are more diverse than this! Another image from Antiquity is that of debate and controversy in the Greek polis.

Students of philosophy remember the sophists, and the dialectical nature of Platonic dialogues, where Socrates corners his opponents by clever moves at the right time. Here is a picture demonstrating this style of reasoning, painted by Rubens (you should count five philosophers: four alive, and one dead). This other view of logic and argumentation is much more like a *game*. There are many players, what they say is in response to each other – and debates can typically have the bitter taste of losing, just because you timed your moves in the wrong order.

Logic and computation Over the centuries, logic also became associated with computation and machines. Leibniz' famous recommendation for resolving logical disputes was "Calculemus". Parties would code their differences of opinion into formulas, after which the right or wrong of the matter could be settled by mere binary arithmetic on code. Note that this again reduces a multi-agent debate to the lonely workings of a computing device. Anyway, computing machines have come about, and to-day's digital computers are the direct descendants of the 'Turing machines' proposed in the 1930s for analyzing the scope and limits of mathematical computation.



Interestingly, one of the earliest key results was a negative one. Some simple questions, such as that whether a given machine will halt on a certain input – or whether your computer will 'freeze' following a particular key stroke of yours – turned out *undecidable*. There just is no guaranteed method finding out the right answer!

Similar limitative results about the scope of formal proof were discovered in Gödel's famous Incompleteness Theorems. Even so, mathematical logic and computer science have thrived in the 20th century, finding all sorts of computation and proof devices that underlie the revolutionary information processing of to-day. Indeed, when TIME Magazine published a list of 'Twenty most influential intellectuals' of the 20th Century, it included Turing, Gödel, and Wittgenstein, the most congenial philosopher to all this.

Back to conversation And yet, the old dialectical picture seems as alive as ever! Drawing a conclusion is just one of many ways of obtaining information. We can also *see*, and often just *ask*! A question plus answer are the simplest multi-agent informational episode, and it, too, has clear logical features. Consider the following scenario:

Q "Is this building the Louvre?"

A "Yes, it is."

We do these things thousands of times in our lives. But notice the subtle information flow. Normally, the questioner ***Q*** indicates that he does not know whether this is the Louvre. But also, by addressing ***A***, he makes it clear he thinks that she might know the answer. We convey information about facts, but also about what we know about other people. Next, when the answer is given, ***A*** does not just convey that this is the Louvre. She now also knows that ***Q*** knows, and ***Q*** knows *that*, and so on to further iterations. In a term used by modern philosophers, linguists, and game theorists, ***Q*** and ***A*** achieve *common knowledge* of the Louvre fact. If you think all this just epicycles, imagine that you have found out my bank code. If you know that I do not know that you know, you will be tempted to empty my account. But if you know that I know that you know, then you will probably stay honest. Our behaviour is kept in place by mutual information...

In daily life, we are quite good at manipulating mixed forms of information.

> Jury members *1, 2, 3* must select candidate *A* or *Q*. Each writes his choice on a slip of paper, and a vote teller sees them all. Now the teller says "There is no consensus". Next *2* shows his slip to *1* without showing 3. *1* sighs he still does not know which candidate was elected. Who knows the outcome of the vote?

This mix of assertions, half-hidden actions, and sighs suffices for *3*, but not the other members! After the teller speaks, everyone knows the vote was *AAQ* or *AQQ*. If *1, 2* had voted alike, *1* would know the outcome after seeing *2*'s paper. As he did not, *3*'s vote is decisive. Everyone can follow this reasoning, so *1, 2* do know that *3* knows.

This interactive reasoning with diverse sources permeates our lives, and we like it so much that we even continue at night, playing parlour games like "Cluedo" with complex moves. Such games are a gold-mine of information and a challenge to logic:

Knowledge and many-agent systems Conversation abandons the loneliness of a single prover, and focuses on groups of agents, just like modern computer science. It is crucial that we maintain information about other people – and even more, groups have special forms of knowledge, not reducible to what separate agents know. Just think of a group finding out things new to all its members by sharing information. Modern *epistemic logics* handle these phenomena by keeping track of assertions

$K_i\phi$ agent i knows that ϕ,
$C_G\phi$ ϕ is common knowledge in group G.

Reasoning about epistemic notions turns out to be as precise as in mathematical logic, and complete axiom systems are known. But there is also the dynamics to be studied!

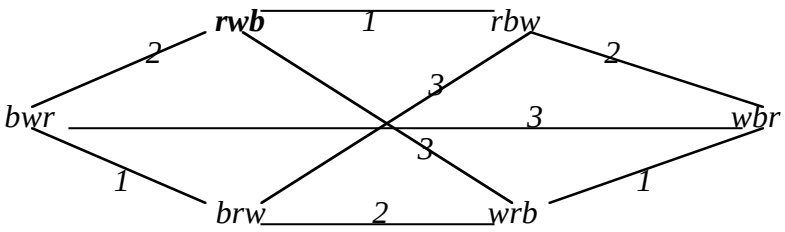
Dynamics of communication Natural language is really a device for cognitive programming. Each successive speech act modifies the current information state of hearers and speakers. We can model this in a simple manner. Consider our Party Puzzle again. At the start, no information was present, and all 8 options remained:

{MAJ, MA-J, M-AJ, M-A-J, -MAJ, -MA-J, -M-AJ, -M-A-J}

Now the three given premises *update* this initial information state, by removing options incompatible with them. In successive steps, we get the following reductions:

(a) ***(M or A) → J*** new state *{MAJ, M-AJ, -MAJ, -M-AJ, -M-A-J}*
(b) ***not-M → A*** new state *{MAJ, M-AJ, -MAJ}*
(c) ***A → not-J*** new state *{M-AJ}*

The same mechanism works in multi-agent settings, such as *card games*. Cards 'red', 'white', 'blue' are dealt to players: *1, 2, 3,* one each. Each sees his own card only. The real distribution over *1, 2, 3* is red, white, blue (***rwb***). We draw the information state:

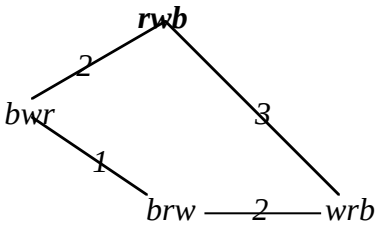


Here, lines encode uncertainty, viz. which deals players find possible. E.g., the *1*-line between *rwb* and *rbw* shows player *1* cannot distinguish these deals, while *2* and *3* can (they have different cards in them). Now the following two moves take place:

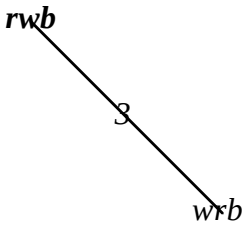
> *2 asks 1 "Do you have the blue card?"*, and *1 answers truthfully "No"*.
> Who knows what then? Here is the effect in words:
>
> Assuming the question is sincere, *2* indicates that she does not know the answer, and so she cannot have the blue card. This tells *1* at once what the deal was. But *3* does 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.

We now give the updates in the diagram, making all these considerations transparent:

After *2*'s question:



After *1*'s answer:



We see at once in the final diagram that *1, 2* know the deal, as they have no uncertainty lines left. But *3* still does not know, but she does know that *1, 2* know – and in fact, this is common knowledge. Similar analyses exist by now for other conversation scenarios, and indeed, for a wide variety of puzzles and games, including "Cluedo".

Logics of programs But conversation involves much more than single assertions. If you need a raise from your boss, you make sure to say the right things in the right order. First praise his inspired leadership, then ask for the money – not the other way round. This is *composition* of programs. Depending on whether he looks upset or relaxed, you choose the right words. This is like an *IF THEN ELSE* in programming. And if one dose of flattery does not work, you keep applying it until it works: this is the crucial *WHILE DO* instruction. Thus, conversational strategies involve all major sequential structures from computer science. And even more sophisticated *parallel* constructions occur, like making students answer your question simultaneously. Thus, not surprisingly, *dynamic logics* of programs developed in computer science since the 1970s have been enlisted for analyzing communication. While these logics originally dealt with numerical programs and the analysis of their behaviour, they now describe any sort of structured action where information flows.

This is one of many recent instances where fundamental ideas from computer science (rather than some practical desktop device) are affecting other disciplines in Academia.

Dynamic-epistemic logic Combining epistemic and dynamic logics gives rise to logical systems with joint assertions which describe effects of communicative actions:

$[!A]K_i\phi$ after a public statement that A is the case, agent i knows that ϕ

By now, complete axiom systems are known, as well as a lot of further detail about the expressive power and complexity of dynamic-epistemic languages. Purely as an illustration, we print the axioms of one such system, which is coming into wide use:

$[!A]p$	$\leftrightarrow$	$A \rightarrow p$	*for atomic facts* p
$[A!]\neg\phi$	$\leftrightarrow$	$A \rightarrow \neg[A!]\phi$	
$[!A](\phi \,\&\, \psi)$	$\leftrightarrow$	$[!A]\phi \,\&\, [!A]\psi$	
$[!A]K_i\phi$	$\leftrightarrow$	$A \rightarrow K_i(A \rightarrow [!A]\phi)$	
$[!A]C_G\phi$	$\leftrightarrow$	$C_G(A, [!A]\,\phi)$	

These axioms analyze complex effects of public assertions A in terms of simpler ones. Such systems for describing multi-agent information flow are as exact and legitimate as any earlier ones. They can deal with many sorts of puzzles, such as the knowledge puzzles for cards that you get for free on some cell phones. When adding known axioms for program operations, they solve famous puzzles going far back into history:

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

In general, with k muddy children, $k–1$ rounds of announcing ignorance occur, after which common knowledge sets in of which children are dirty.

Email, hiding, and lying Muddy Children involves only public statements, out in the open. But actual communication can be much more complex. For instance, in our committee example, showing the slip with your vote to your neighbour, even when done in public view, gives different information to different members. Non-neighbours only see *that* you communicate your vote, not *what* it is. And one step further than this, genuine misleading often takes place. Consider a wonderful new medium like *email*. When you send a message, the button *cc* will make its content common knowledge in your group, as it turns the message into a public announcement. But much finer blends of information and ignorance are achieved by the 'blind carbon copy' button *bcc*, which

sends the message to a subgroup only, unbeknownst to the others. Extended dynamic-epistemic logics have been developed in the past few years which can describe and analyze such more complex forms of communication. They still transform information diagrams, but in much more subtle ways – and these diagrams may even increase the size as a situation gets complicated. This is often the case in games, where mid-play is more complex informationally than the early stages, or the end game.

Yet one more complexity threshold is crossed when we consider *cheating* and *lying*. Many parents think their children speak the truth because of their angelic character. But the real reason is that these subtler logical social skills are only mastered by adults.

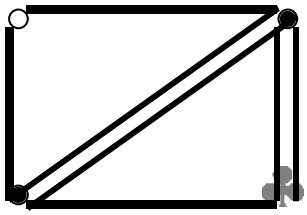
More complexity: conversation planning A further source of complexity in these logic systems is the earlier program structure of conversational strategies. When we turn from just analyzing given scenarios to planning new assertions in such a way that certain desired effects are achieved, the same sort of complexity strikes that was already known for Turing machines. It was shown a few years ago that planning conversation with public announcement and sequential program constructions is an *undecidable* task in general. That does not mean we cannot perform it well – in fact, we do – but it requires creative skill at times. There is no guaranteed automatic method that will get you your raise, or makes you the King or Queen of Conversation.

Belief revision These observations are not the end of logic in its multi-agent dynamic mode, but only the beginning! Beyond updating information, many further cognitive processes play a role in communication. We are often surprised by new observations, or we are contradicted by others, and then we have to *revise our beliefs*. In all we do, we live in a cocoon of expectations about facts, and about others, that guide our thinking and acting – which are constantly modified to keep us attuned to reality. A nice example of such a scenario with information processing agents is this.

One very typical feature of societies of informational agents is their *diversity*. Not everyone has the same knowledge, or processing capacities. We operate under certain assumptions, that may have to be revised. The cult movie *Memento* is a nice example. The protagonist Guy Peirce has lost his long-term memory, making him more like a finite automaton than a Turing Machine. Evil other agents, like Carrie-Ann Moss, learn this to their surprise, and then start taking advantage of his special nature.

Games and strategies Another recent development takes the interaction between different agents seriously in itself. The most pregnant model for such longer-term interaction are *games*. For instance, many communicative settings naturally suggest 'knowledge games' of various sorts. Games also have much to do with logic, as we already pointed out that argumentation itself is a sort of game. It involves a kind of sequential give-and-take, where one's best move depends on preceding ones by others. Here is a simple scenario, which illustrates the response character of playing a game:

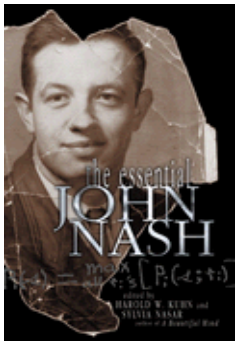
> *Save the Treasure!* Your tribe's favourite Idol is coveted by an American archeologist ***A*** with a long whip, and you must save it. ***A*** is at the white location left in the following diagram, and the Idol at the gray flower:



> Each line gives a possible connection, and the game unfolds as follows. You start by cutting a link, then the archeologist moves along a path, and so on. Can you prevent him from reaching the Idol, or can he always get there?

You may want to 'block' ***A*** straightaway at the white point. But then you will lose! If you cut the upper path, ***A*** moves to the lower black dot, so you must then cut the lower path. But then he moves diagonally upward, and you are too late to cut the two remaining paths toward the Idol. And so on. Still, you do have *a winning strategy*, by first cutting one of the paths to the right, and then keep cutting paths to the Idol depending on where ***A*** moves. If you do this well, time is on your side (just), and you will eventually make the Idol inaccessible to him. (But watch out for that long whip...)

Strategic interaction has been the typical domain of *game theory*. Indeed, one of the oldest results in the area is Zermelo's Theorem (1913), which says, that, in finite games of the sort described here, one of the two players must always have a *winning strategy*, i.e., a guaranteed method for winning against the opponent. Zermelo himself was interested in Chess, where his analysis shows that either White has a winning strategy, or Black has a 'non-losing strategy'. Unfortunately, a century later, we still do not know which – as the full game tree of Chess is stupendously large. Zermelo's theorem was rediscovered by the only world champion in Chess ever produced by our own country: actually, Max Euwe published the same result in 1929. But the real patron saint of game theory is John Nash, whose tragic life inspired the movie "A Beautiful Mind":



Of course, ever since Nash and his predecessors von Neumann and Borel, real game theory also involves *utilities*, and computation of strategic equilibria between players having more refined goals over time than just winning or losing. All this further structure is relevant to understanding multi-agent interaction, and interfaces between fields are broadening. For instance, a new German Heisenberg Centre is devoted to the game-theoretic analysis of *linguistic communication*, and how meanings emerge between speakers and hearers as stable equilibria in communication games. At the opening ceremony, John Nash' co-Nobel Prize winner Reinhard Selten spoke, since his recent interests have turned to language, decision making and cognitive science.

This trend toward placing multi-agent interaction at centre stage can also be observed in modern logic, in studies of such diverse tasks as argumentation, or constructing, or comparing formal mathematical models. Conversely, modern logical systems have also been used to analyze the foundations of game theory. There they provide much richer models for the detailed actions and deliberations of players than the simple 'matrices' from Von Neuman & Morgenstern that may still come to mind to most readers. Indeed, last year's economics Nobel prize winner Robert Auman has been a pioneer in introducing techniques from the above-mentioned epistemic logic into game theory.

In particular, in recent years, part of the emphasis has shifted from finite games that model terminating activities to infinite *evolutionary games*, describing some stable practice in a community in the long run. Such analyses have been applied to predator-prey populations in biology, but also to the evolution of cooperation in human societies. But they apply very well also to the 'operating system' underlying linguistic or logical practices. More technically, this still longer temporal perspective sits well with so-called dynamic and temporal logics of infinite processes in computer science, and game semantics for contemporary 'linear logic'.

Cognitive science Our presentation of current developments has been mostly a priori: mathematical, logical, and computational. But what do people *really* do, in

conversation, communication or games? At the start of the foundational era, such empirical information was deemed irrelevant, and brushed aside by Frege – just at the time when psychology started getting interesting. By now, Frege's Taboo has played itself out. Logicians are getting intrigued by data from cognitive science about actual performance, since these data are not so much an record of human fallacies and follies as an inspiring set of stable and effective practices that demand explanation.

Indeed, intelligent interaction seems a somewhat neglected focus in empirical cognitive science so far. Turning on the just-acquired expensive magnet, one pores into the brains of individual observers, learners, or reasoners, and measures what clicks, boils, and flickers there. But what often remains out of scope is the fact that most intelligent activities involve the interplay of *several agents*: speakers and hearers in a conversation, students and teachers in a classroom, academic researchers in a seminar, and so on. Intelligence is not just an individual, but also a social phenomenon!

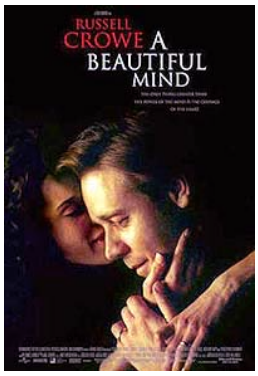
Metaphors and reality The resulting setting for modern logical research is a Triangle. There is of course *logical theory*, there is the *empirical reality* of existing reasoning and updating practices, but intriguingly, there is also a third vertex of *designed new practices*, often virtual, usually inspired by computational ideas. And ideas can flow freely between all three points. For instance, our dynamic-epistemic logics suggest deep analogies between *computation* and *conversation*. In that light, the earlier-mentioned undecidability result for conversation planning really says this:

The computing power of groups of agents equals that of universal computers!

The four philosophers in Rubens' painting rival Turing's machine depicted earlier. Once you realize this, your view of crowded Paris cafés will never be the same again...

But these metaphors do not just help us reinterpret reality, they can also change and enrich it. *Email* was one intriguing new social practice engendered by computation. A more ambitious research program is called *social software*: the design of new social practices with logical-computational methodology. This requires an understanding of algorithms-under-uncertainty of the sort described above. Whether these brave new practices will 'grip', is up to the cognitive scientists again.

Here is a nice example of seamless interplay between old-fashioned and virtual reality. Sylvia Nazar tells that John Nash really appreciated the movie "A Beautiful Mind". When he remarried his estranged wife Alicia some years ago, he managed to kiss her several times at the moment supreme of the ceremony. When asked why, he answered: "I am sure that Russell Crowe also staged several takes of that scene with Jennifer Conolly". The once lonesome Nash had become a multi-agent system:



From secure foundations to dynamic repair It is time to close the circle of this presentation. Modern logic started its great flowering in the quest for absolute certainty, and secure foundations for mathematics. By now, it is clear that no such foundations exist. The true stability and success of our cognitive practices has to do with the dynamic interactive ways in which we process information, and the quality of our adaptive mechanisms for correcting beliefs once they become problematic. Thus, logic is not just the guardian of eternal safety in a world that has been sanitized of all contradictions. It is just as much about how we correct ourselves, mostly in interaction with others. Thus, logic is rather the dynamic and social *immune system of the mind*.

References There is a flourishing community at the interface of logic, linguistics, computer science, artificial intelligence, and game theory, with many conferences like LICS, CSL, TARK, JELIA, LOFT, and since 15 years, the European ESSLLI summer schools (http://www.folli.org). Also informative are two websites on Logic, Games, and Computation at the Institute for Logic, Language and Computation in Amsterdam: http://www.illc.uva.nl/lgc/, http://www.illc.uva.nl/GLoRiClass/, or the Netherlands Institute for Advanced Studies: http://www.nias.knaw.nl/en/research_group_2006_07/nucleus/

Also, here are a few publications leading into the area:

P. Adriaans & J. van Benthem, eds., 2007, *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. Many updated versions.

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

J. van Benthem, 2005A, 'Open Problems in Game Logics', in S. Artemov et al., eds., *Essays in Honour of Dov Gabbay*, King's College Publications, London, 229–264. Available also on http://staff.science.uva.nl/~johan/

B. de Bruin, 2005, *Explaining Games*, Dissertation, Institute for Logic, Language and Computation, Amsterdam, winner Praemium Erasmianum 2006.

H. van Ditmarsch, W. van der Hoek & B. Kooi2007, *Dynamic Epistemic Logic*, Kluwer-Springer Academic Publishers, Dordrecht.

European Science Foundation, 2006, 'Modeling Intelligent Interaction: Logic in the humanities, social and computational sciences', Strassbourg, Eurocores Project LogiCCC, http://www.esf.org/

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

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.

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

R. Muskens, J. van Benthem & A. Visser, 1997, 'Dynamics', a chapter in J. van Benthem & A. ter Meulen, eds., *Handbook of Logic and Language*, Elsevier Science Publishers, Amsterdam, 587-648.

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

R. Stalnaker, 1999, 'Extensive and Strategic Form: Games and Models for Games', *Research in Economics* 53:2, 93-291.