Game Theory, Logic and Rational Choice

Johan van Benthem and Jan van Eijck

A game theorist has joined the group. Our usual protagonists use the occasion to clarify what game theory and logic might have to say about rationality of actions.

Philosopher: What I would like to understand better is how game theory can help us to understand rational choice, and how this is related to logic. Philosophy has a long-standing interest in rational behavior. The hallmark of rationality is always taken to be the "good reasons for acting" that rational people can give. But game theory seems more concerned with what people actually do than in the reasons they might care to give for their actions.

Game Theorist: In game theory, it is common practice to analyze problems of rational choice as problems of finding the best move in a game. A player is rational if she plays according to an optimal strategy. Game theory has various ways of determining whether a given strategy is rational. For finite games of complete information the preferred method is backward induction. I take it that you all know how that works.

Philosopher: Yes, yes. But I suppose it will do us no harm if you briefly remind us.

Game Theorist: Backward induction is a technique to solve a finite game of perfect information. First, one determines the optimal strategy of the player who makes the last move in the game. Then, taking these moves as given future actions, one determines the optimal strategy for the next-to-last player in the game. And so on, backwards in time, until the beginning of the game is reached. As it turns out this determines the Nash equilibrium of each subgame of the game.

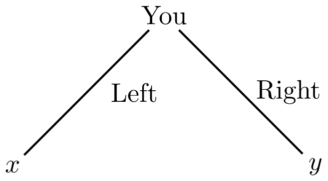
Philosopher: Splendid. And what, again, if I may ask, are Nash equilibria and subgames? I have heard of John Nash, of course. I enjoyed watching *A beautiful mind.*

Game Theorist: The movie is certainly entertaining, but the book [9] on which it is based gives a more accurate picture of the life of John Nash. But I see you want further memory refreshment. A Nash equilibrium, also called strategic equilibrium, is a list of strategies, one for each player, which has the property that no player can achieve a better payoff by unilaterally changing her strategy. A subgame is a piece of a sequential game beginning at some node such that each player knows every action of the players that moved before him at every point. There are excellent textbooks where further details can be found. I particularly recommend [12] and [10]. And [8] collects all contributions to game theory by John Nash, with an enlightening introduction.

Logician: Of course it is also possible to take games as objects in their own right, and study transformations on them. Instead of condemning a particular move as irrational, one might wish to take the move as a revelation of an agents preference. This transforms a given game into a new one, with different preferences. Also, maybe a move reveals an agent's belief about the beliefs of the other game participants. This would correspond with a game transformation where beliefs change. Still a different way of changing a game is by making a promise: This changes other agents' expectations, so it also corresponds to a game transformation.

Computer Scientist: This smells of the update operations of dynamic logic.

Logician: You are quite right. A more general study of game transformations would involve dynamic and epistemic game logics. But I propose not to dive into that, but instead to look at the structure of very simple choices. Suppose you have a choice between two available actions *Left* and *Right.* *(Writes on the whiteboard.)*



The choice is yours. What will you do?

Game Theorist: Well, I suppose that without further information no prediction can be made. A game theorist would say that we need to know the values you attach to the outcomes x and y. Or stated in another way, your preferences between these.

Computer Scientist: It looks to me that the logical form of the prediction is this:

> You must (and can) do *Left* or *Right*.
> You prefer outcome x.
> Therefore: You will perform action *Left*.

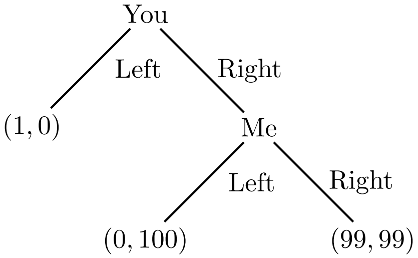
Logician: Surely, there is no compelling logical reason why you must do what is best for you. Much of the greatest world literature is about people who do not. But one might say that rational people behave according to this inference pattern, and hence we could take it as a definition of behavior for a certain kind of agent.

Computer Scientist: I suppose the pattern of inference could also be invoked post-hoc. When I want to explain the way you behave when you choose *Left*, I conclude that you must have liked outcome x better than outcome y.

Philosopher: We do this all the time when "rationalizing" our own actions to ourselves or others. You chose action *Left* without thinking about the consequences in the cosy half-dark of a late night bar—but in the harsh light of the next morning, waking up with a headache in some unknown place, you have no shortage of good reasons for your behavior.

Computer Scientist: Yes, we humans may not be very good in taking rational decisions with a strict logical discipline beforehand, but we are wizards in rationalizing our actions afterwards.

Logician: Moaning over human nature gives great satisfaction, doesn't it? Snap out of it, guys. I have something more interesting for you to look at, a case where two agents interact. Let's assume payoffs are also given, to represent the agents' preferences. You first choose *Left* or *Right*. If you choose *Left* the game is over; while if you choose *Right*, it is then my turn to choose between *Left* and *Right*. The payoffs are indicated in the following game tree, with your value written first, then mine. *(Writes on the whiteboard.)*



Game Theorist: OK. The standard procedure in game theory for this scenario is BI (Backward Induction). We start at the bottom: as a "rational" player, I will choose to go *Left*, since 100 is better than 99. You can see this coming: so going *Right* gives you only 0, whereas going *Left* gives you 1. Therefore, you will choose *Left* at the start, and we both end up getting very little, while I lose most of all.

Computer Scientist: Ahem, rationality seems to come at a rather high price in this example.

Game Theorist: Much more sophisticated scenarios exist where standard game solution procedures have strange effects.

Logician: Let's not quibble about whether this is right or wrong. Let us look instead at what the example can teach us about the logical underpinnings of BI. What we see is interaction between agents, where your expectations about my behavior determine the outcome. In particular, you assume that I am rational in the game theoretical sense, choosing *Left*, predicting that *Right* will end in $(0, 100)$. And so on, in more complex games. BI is often considered the "standard solution procedure" for games. But what is the status of this mixture of available actions, preferences, and expectations?

Game Theorist: BI-style rationality has a remarkable staying power. It may not be a great predictor of human behavior, but it has its use for rationally reconstructing it. And, as was remarked earlier, there seems to be a universal need for such rationalizations.

Philosopher: But if I assume that your preferences only reveal themselves in how you play, then your rationality becomes a truism. Suppose that your preferences between the outcomes of some given game are not known. Then I can always ascribe preferences to you which make your actions rational in the BI sense. In the simplest scenario, if you choose action *Left* over *Right*, I can always make your given choice appear rational *a posteriori*, by assuming that you prefer the former outcome over the latter.

Game Theorist: Yes, but let us pursue this. This style of rationalization carries over to more complex interactive settings. For now one must also think about me, i.e., the other player that you are interacting with. Let a finite two-player extensive game **G** specify my preferences, but not yours. Moreover, let both our strategies $\sigma_{\mathrm{me}}, \sigma_{\mathrm{you}}$ for playing **G** be fixed in advance, yielding an expanded structure that is sometimes called a "game model" **M**. Now, here is a technical question. When can we rationalize your given behavior σ_{you} to make our two strategies the BI solution of the game?

Logician: In principle, to achieve this, we have complete freedom to just set your preferences, or equivalently, set the values which you attach to outcomes of the game. And this can be done independently from my already given evaluation of these outcomes.

Game Theorist: Even so, not all game models **M** support BI. In particular, my given actions encoded in σ_{me} must have a certain quality to begin with, related to my given preferences. Note that, at any node where I must move, playing on according to our two given strategies already fixes a unique outcome of the game. What is clearly necessary for any successful BI-style analysis, then, is this. My strategy chooses a move leading to an outcome which is at least as good for me as any other outcome that might arise by choosing an action, and then continuing with $\sigma_{\mathrm{me}}, \sigma_{\mathrm{you}}$.

Logician: There is a folklore result about such games that are "best-responsive" for me.

> In any game that is best-responsive for me, there exists a preference relation for you among outcomes making the unique path that plays our given strategies against each other the BI solution.

To see why this is true, start with final choices for players near the bottom of the game tree, assigning values reflecting preferences for you as described before. Now proceed inductively. At my turns higher up in the game tree, their being best-responsive for me ensures automatically that I am doing the right thing, provided our strategies in the subgames following my available moves are already in accordance with BI. Next, suppose it is your turn, while the same inductive assumption holds about the immediate subgames. In particular, then, these subgames already have BI-values for both you and me.

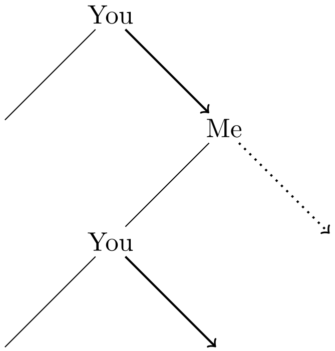
Now suppose your given move a in σ_{you} leads to a subgame which has a lower value for you than some subgame produced by another move of yours. In that case, a simple trick makes a the best for you. Take some fixed number N large enough so that adding it to all outcomes in the subtree headed by a makes them better than all outcomes reachable by your other moves than a. Now, it is easy to see the following feature:

> Raising all your values of outcomes in a game tree by a fixed amount N does not change the BI-solution, though it raises your total value by N.

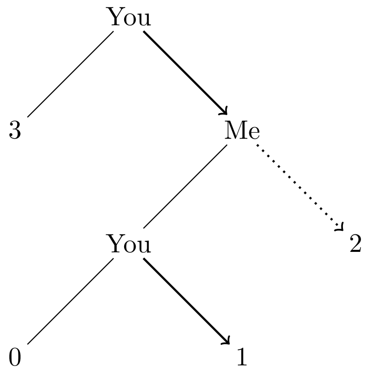
So doing this to a's subtree, your given move at this turn has become best.

Philosopher: An example would do me no harm at this stage.

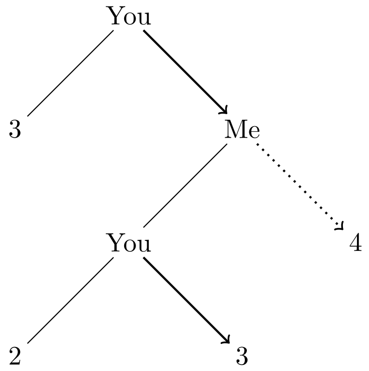
Logician: At your service. Here is a picture of an example game between you and me, with solid arrows for your given moves and dotted arrows for mine. No payoffs are indicated, for it is assumed that your payoffs are not known.



Let us fill in payoffs for you that make it appear that your behaviour in the game was irrational.



Assume that the value 3 on the left has been assigned in some subgame already. Now adjust the values in the subgame that results from your first move. An adjustment that works is adding 2 to all of them. Of course, the adjustment can be made in many ways to get BI right.



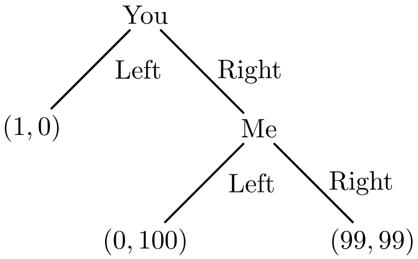
Game Theorist: So the conclusion must be that one can always pretend that you did the rational thing by tinkering *post facto* with your preferences. This is the basis for re-analysis of games in practice, replacing initial assignments of values for players by others so as to match observed behavior.

Logician: But there are alternative ways of rationalizing observed behavior. What we have seen so far takes the strategies, with their accompanying beliefs, as given, and uses these to work out the preferences for one of the players.

But one could also start from given preferences for both players, and use these to modify the beliefs of the players to rationalize the given behavior.

Philosopher: A simple example again, if you please.

Logician: Look again at our very first example. *(Points at the whiteboard.)*



Suppose you choose *Right* in this game. One can interpret this rationally if we assume that you believe that I will go *Right* as well in the next move. This rationalization is not in terms of your preferences, but of your beliefs about me.

Game Theorist: This style of rationalizing need not produce the BI solution.

Logician: No, but it still presupposes a certain pattern in a game **G**, or better, game model **M**. This time, consider a finite extensive game as before, with your strategy σ_{you} and your preference relation given. My preference relation does not matter in this scenario.

Game Theorist: Not all behavior of yours can be rationalized in this way. For suppose that you have a choice between two moves *Left* and *Right*, but all outcomes of *Left* are better than all those arising after *Right.* Then no beliefs of yours about my subsequent moves can make a choice for *Right* come out "best".

Logician: To put it differently, a game model which can be expanded so as to make your moves best in terms of your beliefs about my strategy must satisfy the following condition:

> Your strategy σ_{you} never prescribes a move for which each outcome reachable via further play according to σ_{you} and any moves of mine is worse than all outcomes reachable via some other move for me.

Game Theorist: That's right. In case you are the last to move, this coincides with the usual decision-theoretic requirement that you must choose a move that guarantees a best possible outcome for you.

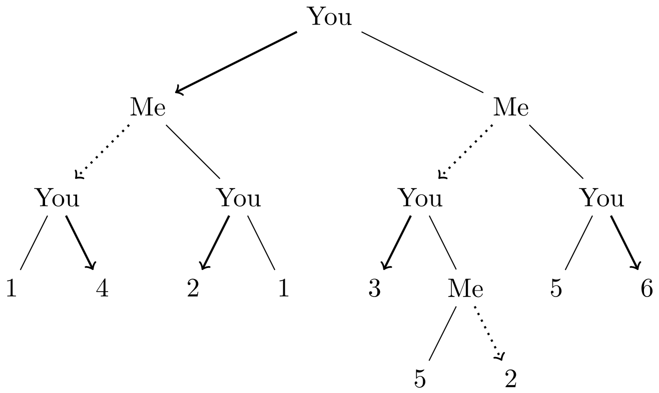
Logician: Let us call a game model satisfying this condition "not-too-bad" for you. [4] has the following theorem: In any game that is not-too-bad for you, there exists a strategy τ for me against which, if you believe that I will play τ against your σ_{you}, is optimal. Why is this true? This time, the adjustment procedure for finding the rationalizing strategy is a bit different. The idea works *top-down* along the given game tree. Suppose that you make a move a right now according to your strategy. Since your given strategy σ_{you} is not-too-bad for you, each alternative move b of yours must have at least one reachable outcome y (via σ_{you} plus some suitable sequence of moves for me) which is majorized by some reachable outcome x via a. In particular, the *maximum outcome value* for you reachable by playing a will always be better than some value in the subgame for the other moves.

Game Theorist: You still have to explain why your given move a is optimal.

Logician: Right. Here is the expected strategy for *me* which makes it optimal. Choose later moves for me in the subgame for a which lead to the outcome x, and choose moves for me leading to outcomes $y \leq x$ in the subgames for my other moves b. Doing this makes sure a is a best response against any strategy of mine that includes those moves. This does not yet fully determine the strategy that you believe I will play, but one can proceed downward along the given game tree. *(To the philosopher.)* And now I suppose you want to see an example again?

Philosopher: Yes, if you don't mind.

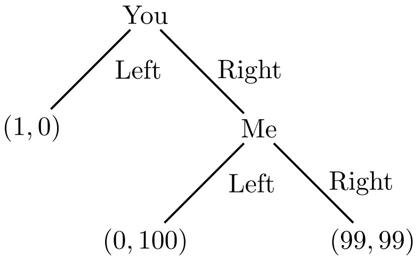
Logician: Not at all. Here is a game with your moves marked as solid arrows, and with the necessary rationalized beliefs about me indicated by the dotted arrows. Note that in contrast with the folklore result I mentioned before, the outcome values for you are now given beforehand. *(Writes on the whiteboard.)*



Your initial choice for going *Left* has been rationalized by forcing the outcome 4–assuming that I will go *Left*—which is better than the forced outcome 3 on the right—assuming that I would go *Left* there, too. Likewise, one step further down, in the subtree with outcomes 3, 5, 2, a *Right* move for you would have resulted in 2 rather than 3, if we assume that I would next go *Right* there.

Philosopher: I see.

Logician: Mind you, the theorem provides no underpinning of your belief that I will play τ. Indeed, τ may go totally against my known preferences. But the rationalization becomes more convincing if we can think up some plausible story of why I might want to act according to τ. And this is sometimes possible in ways different from BI. Look once again at our earlier example. *(Points at the whiteboard.)*



Think of why I might believe that you will choose *Right* in this game. Here is a plausible story. If a player has run risks for the "common good" by doing the other player a favor, he should not be punished for that, but rewarded, the argument goes. In this particular example, I run the risk of losing one point in playing *Right.* Hence you owe me at least that much—and you should reward me by choosing an outcome where I do not lose that point. This story is worked out in [3], by the way, where a candidate for a general alternative to BI is put forward in terms of *Returning Favors.*

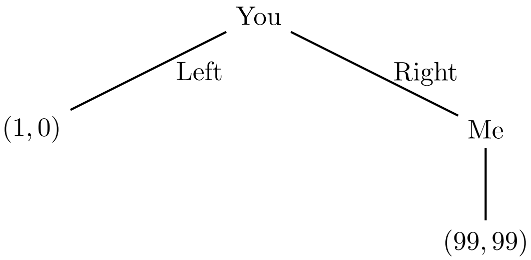
Philosopher: Maybe we should then look at *BI-style reanalysis* and *Returning Favors-style reanalysis* as two different ways of making sense of the same behavior? Surely these are just extreme cases of rationalizing given strategies in games.

Logician: Yes, indeed. And I suppose the moral is that we could devise procedures manipulating both my preferences and beliefs. But instead of rationalizing what has happened already, we can also try to do something about the initial situation we find ourselves in. Look at our running example again. What could I possibly do to break out of the scenario we are in, and change it in my favor?

Philosopher: Well, I suppose you could make a promise to the other player. Or rather, "I" could make a promise to "You", if you see what I mean. "I herewith solemnly promise that I will not go *Left* when you have gone *Right.*" That should do the trick, if we suppose you know that I am honest. I mean, if we suppose "You" knows that "I" is honest. Well, you know what I mean.

Logician: Do you mean " 'You' knows what 'I' means"? But I shouldn't tease you, for you are quite right.

Game Theorist: As the KGB officer said, we can always force people to be honest. Let us say that my promise puts such a high punishment on my choosing *Left* that this branch disappears from the game tree. This would give the following new game:



I suppose the general question becomes how to model a process where games can change because of certain actions?

Computer Scientist: Dynamic logics of information update. I knew it. I saw it coming.

Logician: How perceptive you are! A binding promise is like a public announcement $!\phi$ of a true assertion ϕ. To be precise here, we should work again with game models $\mathbf{M}$, not just games. Details are given in [1; 2].

Computer Scientist: A public announcement ϕ restricts the current $\mathbf{M}, s$ to a model $\mathbf{M}|\phi, s$ of just those worlds in $\mathbf{M}$ which satisfy ϕ. One can then analyze effects of making announcements on agents' beliefs in dynamic epistemic or doxastic logics which involve valid "reduction axioms" such as:

$$[!\phi]B_i\psi \leftrightarrow \phi \rightarrow B_i(\phi, [!\phi]\psi).$$

Here $[!\phi]\psi$ expresses that ψ holds after public announcement ϕ, and $B_i(\_, \_)$ is used for conditional belief. Note that the axiom pushes the $[!\phi]$ operator past the belief operator.

Logician: In our game scenario, a promise announces an intention in a game, which restricts the possible reachable nodes. For a complete logic for game-changing by promises and announced intentions, one needs a language over game models which describes players' moves, preferences, and beliefs. A good test on whether the right expressive power has been achieved is definability of the BI solution. There is already an extensive literature on this: [1; 6; 7; 5]. Never mind the details.

Game Theorist: I suppose that if you make the base logic strong enough you can prove completeness by reduction, just as in the case of public announcement logic. Let me guess:

> There is a complete logic of public announcements over extensive games of perfect information which consist of a standard static base logic plus a complete set of reduction axioms for announcement modalities over the relevant move and preference modalities of the game language.

Am I right?

Logician: Yes, but maybe the more interesting issue concerning behavior is how public announcement of intentions changes what we know about the effects of strategies in a game. Strategies can be defined as programs in a dynamic logic over extensive games [2], which can then define a modality $\{\sigma\}\phi$ saying that strategy σ only leads to nodes satisfying condition ϕ. Now we can also give reduction axioms for reasoning about the effects of strategies in the changed game.

Computer Scientist: Reduction axioms again. Push the promise through the strategy operator, and you are done. So the axiom for the changing power of a promise A should be something like "$[!\phi]\{\sigma\}\psi$ is equivalent to $\{\sigma\}[!\phi]\psi$." Am I right?

Logician: Absolutely right. This all uses good old propositional dynamic logic [11]. The result uses the insight that propositional dynamic logic is closed under domain relativization. It applies to reasoning about the new BI-strategies in our earlier games changed by a promise.

Philosopher: I guess this gets us at procedural conceptions of rationality, as following the right procedure to improve one's situation. But how about issuing threats, not just promises?

Computer Scientist: We can handle that by substitution. Just a matter of replacing carrots by sticks. That shouldn't cause any technical difficulties. There may be ethical implications, but we feel we can safely leave such matters with you.

(They all smile at the philosopher.)

References

[1] J. van Benthem. Games in dynamic epistemic logic. *Bulletin of Economic Research*, 53(4):219–248, 2001.

[2] J. van Benthem. Extensive games as process models. *Journal of Logic, Language and Information*, 11:289–313, 2002.

[3] J. van Benthem. Rational dynamics and epistemic logic in games. In S. Vannucci, editor, *Logic, Game Theory and Social Choice III*, pages 19–23, 2003. To appear in *International Journal of Game Theory.*

[4] J. van Benthem. Rationalisations and promises in games. Handout, University of Beijing, October 2006.

[5] J. van Benthem, S. van Otterloo, and O. Roy. Preference logic, conditionals, and solution concepts in games. In H. Lagerlund, S. Lindström, and R. Sliwinski, editors, *Modality Matters*, pages 61–76, 2006.

[6] B. de Bruin. *Explaining Games.* PhD thesis, ILLC, Amsterdam, 2004.

[7] P. Harrenstein. *Logic in Conflict.* PhD thesis, Department of Computer Science, University of Utrecht, 2004.

[8] Harold W. Kuhn and Sylvia Nasar, editors. *The Essential John Nash.* Princeton University Press, Princeton and Oxford, 2002.

[9] Sylvia Nasar. *A Beautiful Mind.* Simon and Schuster, New York, 1998.

[10] Martin J. Osborne. *An Introduction to Game Theory.* Oxford University Press, 2004.

[11] V. Pratt. Semantical considerations on Floyd–Hoare logic. *Proceedings 17th IEEE Symposium on Foundations of Computer Science*, pages 109–121, 1976.

[12] Philip D. Straffin. *Game Theory and Strategy.* The Mathematical Association of America, New Mathematical Library, 1993. Fourth printing: 2002.