Rationalizations and Promises in Games

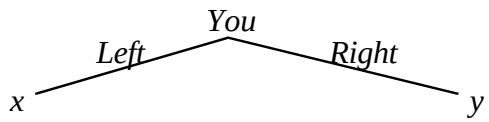
Johan van Benthem, Amsterdam & Stanford

1 October 2006

Abstract Understanding human behaviour involves "why"'s as well as "how"'s. Rational people have good reasons for acting, but it can be hard to find out what these were and how they worked. In this Note, we discuss a few ways in which actions, preferences, and expectations are intermingled. This mixture is especially clear with the well-known solution procedure for extensive games called 'Backward Induction'. In particular, we discuss three scenarios for analyzing behaviour in a game. One can rationalize given moves as *revealing agents' preferences*, one can also rationalize them as *revealing agents' beliefs* about others, but one can also *change* a predicted pattern of behaviour by *making promises*. All three scenarios transform given games to new ones, and we prove some results about their scope. A more general view of relevant game transformations would involve dynamic and epistemic game logics. Finally, our analysis describes and disentangles matters: but it will not tell you what to do!

1 Reasons for actions

You can perform one of two available actions *Left* and *Right*:



The choice is yours. What will you do? Without further information, no prediction can be made. Philosophers and decision theorists say that we need to know the values you attach to the outcomes *x*, *y* – or stated in another way, your *preferences* between these. Then, the logical form of the prediction is often said to be this:

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

But 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 behaviour for a certain kind of agent.

The same pattern of inference is often invoked post-hoc, when we explain observed behaviour. I see you choose *Left*, and conclude you must have liked outcome *x* better

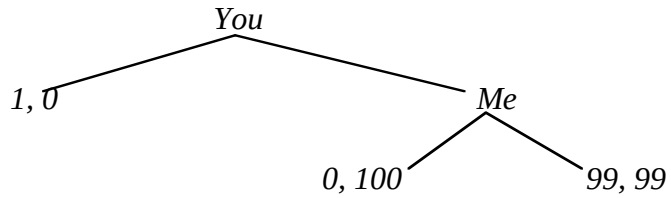
than *y*. We even do this when 'rationalizing' our own actions to ourselves or others. You chose action *Left* without thinking about the consequences in a shady mid-night bar – but oh, those good reasons you can add in the following morning light! We humans may not be very good in taking rational decisions with a strict logical discipline beforehand, but we are wizzards in rationalizing our actions afterwards.

The entanglement of actions, preferences, and rationality gets even more complicated in the context of *games*, where more than one agent is involved. Let usgo there.

2 Interactive rationality in games

Here is a simple interactive scenario. You and I are about to start a little game:

Example 1 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 pay-offs are indicated in the following game tree, with your value written first, then mine:



The standard procedure in game theory for this scenario is '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. Rationality literally has a high price!

Much more sophisticated scenarios exist where standard game solution procedures have strange effects. My concern in this Note is not that this is 'wrong'. Underneath various veneers, many human interactions work on 'using' and 'being used', and rational suspicion is a fact of life. But my interest is in the logical reasoning underpinning Backward Induction. This is more complex than what we have with single decisions, since it also involves a new feature, viz. your *expectations* about my behaviour. In particular, you assume that I am rational in the above sense, choosing *Left*, predicting that *Right* will end in 0, 100 – and so on, in more complex games. Backward Induction is often considered the 'standard solution procedure' for games. What is the status of this mixture of available actions, preferences, and expectations? The following sections present three ways of viewing the above pattern of reasoning.

3 Given actions and revealed preferences

Rationality in the above sense of decision theory and backward induction has a remarkable staying power. One reason for this is its role, not so much in predicting human behaviour, but in *rationally reconstructing* it, the earlier-mentioned process of 'rationalization'. Suppose that your preferences between the outcomes of some given game are not known. Then one can always ascribe preferences to you which make your actions rational in the above sense. In the simplest scenario, if you choose action *Left* over *Right*, one can always make your given choice rational a posteriori by assuming that you prefer the former outcome over the latter.

This style of rationalization carries over to more complex interactive settings. But now one must also think about me, i.e., the other player that you are interacting with. Let a finite two-player extensive game $\boldsymbol{G}$ specify my preferences, but not yours. Moreover, let both our strategies σ_{me}, σ_{you} for playing $\boldsymbol{G}$ be fixed in advance, yielding an expanded structure that is sometimes called a 'game model' $\boldsymbol{M}$. Now, when can we rationalize your given behaviour σ_{you} to *make* our two strategies the Backward Induction solution ('*BI*', for short) solution of the game? 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.

Even so, not all game models $\boldsymbol{M}$ support Backward Induction. 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 σ_{me}, σ_{you}.

Let us call such a game *'best-responsive'* for me. The following result is folklore:

Theorem 1 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 Backward Induction solution.

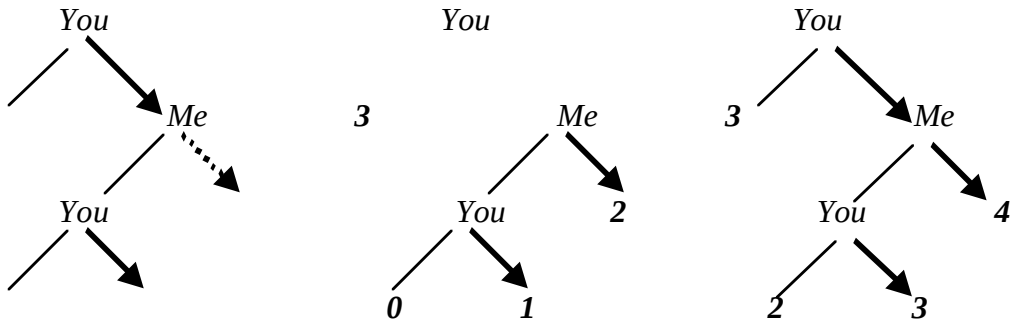
Proof Sketch One starts with final choices for players near the bottom of the game tree, assigning values reflecting preferences for you as described at the beginning of this section. Now proceed inductively. At *my* turns higher up in the game tree, their being best-responsive for me makes sure automatically 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 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*.

Doing this to *a*'s subtree, your given move at this turn has become best. ♣

Example 2 Here is a picture of how our procedure runs *bottom-up* – with bold face arrows drawn for your given moves, and dotted arrows for mine, while the bold-face numbers at the leaves indicate the values for outcomes that we postulate for you:



(a) *(b)* *(c) [N=2]*

Think of the value ***3*** to the left in *(b)* as having been assigned in some subgame already. Of course, the numbers can be assigned in many ways to get *BI* right.

By Theorem 1, one can always pretend that you did the rational thing by tinkering 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 behaviour. But there are alternative ways of rationalizing observed behaviour!

4 Given actions and revealed beliefs

The preceding analysis fixes the use of *BI* to rationalize given strategies in a game ***G***, and their accompanying beliefs, but it postulates the preferences for one of the players. But there are other ways to 'twist the parameters'! At the opposite extreme, one could start from given preferences for both players, but then *modify the beliefs* of the players to rationalize the given behaviour Again, here is a simple example.

Example 3 Suppose you choose *Right* in the game of Example 1. 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. Note that this style of rationalizing need not produce the *BI* solution.

Again, this rationalization for given strategies presupposes a certain pattern in a game $\boldsymbol{G}$, or better: game model $\boldsymbol{M}$. 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). Evidently, not all behaviour of yours can be rationalized. 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'. More precisely, a game model which can be expanded so as to make your moves bets 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.

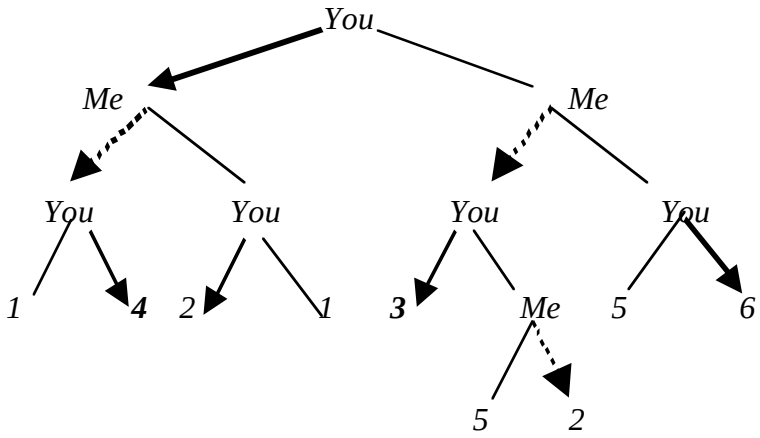
In case you are the last to move, this coincides with the usual decision-theoretic requirement that you must choose a move that guarantees a best possible outcome for you. Let us call a game model satisfying this condition *'not-too-bad'* for you.

Theorem 2 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, your σ_{you} is optimal.

Proof Sketch This time, the adjustment procedure for finding the rationalizing strategy τ 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 reacheable by playing *a* will always be better than some value in the subgame for the other moves.

Now we describe the expected strategy for *me* which makes your given move *a* 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. ♣

Example 4 Here is a game with your moves marked as bold-face arrows, and with the necessary rationalized beliefs about me indicated by the dotted arrows. Note that in contrast with Theorem 1, the outcome values for you are now given beforehand:



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.

Theorem 2 provides no underpinning of your belief that I will play τ. Indeed, τ may go totally against your known preferences! But the rationalization becomes more convincing, of course, 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 Backward Induction. E.g., why might I believe that you will choose *Right* in the game of Examples 1, 3? Van Benthem 2003 suggests a general alternative to *BI* in terms of *Returning Favours*. If players have run risks for the 'common good', they should not be punished for that, but rewarded. In particular, in the given game, 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 it.

Even so, Theorem 1 still applies. Even 'historical justice' can be reanalyzed in the *BI*-style of Section 3, as changing the values which players would attach to outcomes. This is no contradiction: just different ways of making sense of the same behaviour. Also, Theorems 1 and 2 describe extreme cases of rationalizing given strategies in games. We could devise procedures manipulating both my preferences and beliefs.

5 Promises, game change, and dynamic logic

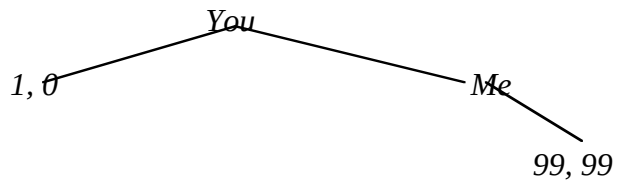
But instead of rationalizing whatever has happened already, we can also try to *do something about* the initial situation we find ourselves in. Clearly, it would be good

for me in the game of Section 2 to change the outcome. And indeed, our life is full of dynamic actions that break out of some given scenario, and *change* it! A good way of changing the world in this way is making a *promise* to the other player:

Example 5 In the game of Example 1, I can promise you that

"I will not go Left when you have gone Right."

In that case, we can both be assured of getting an outcome of 99, as opposed to that meagre outcome 1, 0 recommended by *BI*. Now there are many tricky aspects to this, such as what would convince you that I would *keep* that promise. But these are not my concern: I am just interested in modeling the action in its simplest form. Let us say that my promise puts such a high punishment on my choosing Left that this branch disappears from the game tree. In that case, the new game becomes



How can we model a process where games can *change* because of certain actions? Think of dynamic logics of information update. A binding promise is like

a public announcement *!A* of a true assertion *A*

To be precise here, we should work again with *game models* ***M***, not just games (cf. van Benthem 2001, 2002 for details). A public announcement restricts the current model ***M***, *s* to a model ***M***|*A*, *s* of just those worlds in ***M*** which satisfy *A*. One can then analyze effects of making announcements on agents' beliefs in dynamic epistemic or doxastic logics which involve valid 'reduction axioms'such as:

$$[!A]B_i\phi \leftrightarrow (A \rightarrow B_i([!A]\phi \mid A))$$

where the binary belief operator $B_i(- \mid -)$ on the right stands for a conditional belief. In our game scenario, a promise announces an intention in a game, which restricts the possible reachable nodes. We will merely sketch how this can be done more systematically, referring to the incipient technical literature in the field.

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 Backward Induction solution. Various answers in the literature can be used here: cf. van Benthem 2001, De Bruin 2004, Harrenstein 2004, or van Benthem, van Otterloo & Roy 2006. But once this has been set up, we have this

Theorem 3 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.

As an illustration, here is the key axiom for making a move *a*:

$[!A][a]\phi \leftrightarrow Poss(A) \rightarrow [a][!A]\phi$,
where *Poss(A)* says that at least one reachable end node satisfies *A*.

The required compositional reduction axioms for preference modalities are like those given in van Benthem & Liu 2005, and those for belief are as in van Benthem 2006. We do not repeat them here, as they are well-known and accessible. But maybe the more interesting issue concerning behaviour 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 (van Benthem 2002), 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. These become equivalences of the form

$[!A]\{\sigma\}\phi \leftrightarrow \{\sigma|A\}][!A]\phi$,

using the fact that propositional dynamic logic is closed under domain relativization (van Benthem 1999). This applies to reasoning about the new *BI*-strategies in our earlier games changed by a promise. Eventually, defining equilibria in games requires modal fixed-point logics such as the μ–calculus (van Benthem 2003), in which case we need to exploit the closure of such logics under relativization, and other model-changing constructions (cf. van Benthem, van Eijck & Kooi 2005).

This dynamic take on changing games is more in line with *procedural* conceptions of *rationality*, as following the right procedure to improve one's situation.

6 Where to go from here

Our simple observations raise many further questions. What minimal number of promises will make a given game end in the way I want? What about issuing threats, not just promises? What if players do not know each others' preferences? What about alternatives to Backward Induction, such as Repaying Favours, or Wishful Thinking, assuming that things will always go for the best? What if we do not know other players' strategies, and our beliefs do not fix a single hypothesis? As the game unfolds, we will learn more – but this requires much richer game models than the

ones used here. And what if we recast the scenario of Section 5 as a *new game* over a given one, where making promises itself becomes viewed as an admissible move?

Despite all these open ends, we hope to have shown that a dynamic look at *game transformations* is rewarding, and that it leads to interesting logical questions. Moreover, we think all this helps getting a better grasp of *rationality*, which involves the circular, but intriguing requirement that we be able to interact successfully with other rational agents. In traditional mathematical logic, it was enough to record the structure of 'agent-free' proofs, semantic structures, and valid inferences. But a modern logic for analyzing human behaviour must also have a story of the 'hidden variables': the beliefs, preferences, and the other features which make us tick.

7 References

J. van Benthem, 1999, 'Update as Relativization', Manuscript, ILLC Amsterdam.

J. van Benthem, 2001, 'Games in Dynamic Epistemic Logic', *Bulletin of Economic Research* 53:4, 219–248 (Proceedings LOFT-4, Torino 2000).

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

J. van Benthem, 2003, 'Rational Dynamics and Epistemic Logic in Games, in S. Vannucci, ed., *Logic, Game Theory and Social Choice III*, University of Siena, Department of political economy, 19–23. To appear in *International Journal of Game Theory*.

J. van Benthem, 2006, 'Dynamic Logic of Belief Revision', ILLC Tech Report, Amsterdam. To appear in the *Journal of Applied Non-Classical Logics*.

J. van Benthem, J. van Eijck & B. Kooi, 2005, 'A Logic for Communication and Change', in R. van der Meijden, ed., Proceedings TARK Singapore. Extended version to appear in *Information and Computation*.

J. van Benthem & F. Liu, 2004, 'Diversity of Logical Agents in Games', *Philosophia Scientiae 8:2*, 163–178.

J. van Benthem & F. Liu, 2005, 'Dynamic Logic of Preference Upgrade', to appear in *Journal of Applied Non-Classical Logics*.

J. van Benthem, S. van Otterloo & O. Roy, 2006, 'Preference Logic, Conditionals, and Solution Concepts in Games', in H. Lagerlund, S. Lindström & R. Sliwinski, eds., *Modality Matters*, University of Uppsala, 61 - 76.

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

P. Harrenstein, 2004, *Logic in Conflict*, Dissertation, Institute of Computer Science, University of Utrecht.

S. van Otterloo, 2006, *A Security Analysis of Multi-Agent Protocols*, Dissertation, University of Liverpool and ILLC Amsterdam.