JOHAN VAN BENTHEM[1]

LOGIC IN A SOCIAL SETTING

ABSTRACT

Taking Backward Induction as its running example, this paper explores avenues for a logic of information-driven social action. We use recent results on limit phenomena in knowledge updating and belief revision, procedural rationality, and a 'Theory of Play' analyzing how games are played by different agents.

1 INTRODUCTION

Social agency with many actors has become a major research topic shared between philosophy, computer science, linguistics, and cognitive science. In this paper, my focus will be on *games*, and in particular, the well-known solution method of Backward Induction. Pulling at this one thread brings to light many issues that seem important to a logic of social action *in statu nascendi*. My aim is not to present new technical results in the logic of games, but to explain the guiding questions, and to identify open problems with a broader thrust. In doing so, my aim is to put my preoccupations and presuppositions up for philosophical scrutiny.[2]

2 ENTANGLING ACTION, PREFERENCE, AND BELIEF: BACKWARD INDUCTION

How should we reason about, or inside,[3] social settings involving several agents at once? Social action entangles three basic notions that logicians have tended to study separately: namely, *action*, *belief*, and *preference*. And games are a vivid concrete model for this.[4]

Backward induction All these notions are involved in the standard solution method for extensive two-player games called *Backward Induction* (*BI*). This algorithm computes numerical values for players ***A***, ***E*** at nodes, working bottom

Episteme 8.3 (2011): 227–247
DOI: 10.3366/epi.2011.0019
© Edinburgh University Press
www.eupjournals.com/epi

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press

up according to the following rule:

> *Suppose player* **E** *is to move, and all values for daughter nodes are known. The* **E***-value is the maximum of all the* **E***-values on the daughters, the value for player* **A** *is the minimum of the* **A***-values marked at all* **E***-best daughters. The dual stipulation for the values at* **A***'s turns is completely analogous.*

In the simple game in figure 1, Backward Induction computes a strategic equilibrium where ***A*** goes left at the start, and ***E*** after that (in a hypothetical state that of course never gets reached when ***A*** goes left), with outcomes *(1, 0)* for the players ***A***, ***E***, respectively.

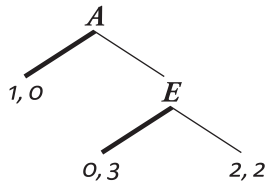


Figure 1

This method produces a *strategy* (let's call it *bi*) for both players by moving from a node for the player whose turn it is to any successor node with a maximal value for her.[5]

Strategies as relations In general, people navigate through social situations with a plan of action that need not be a fully determined rule. Hence, we will generalize strategies from functions to relations merely narrowing down the set of all available moves to one's 'best actions'. Technically, strategies then become binary subrelations of the total *move* relation that can be studied by known logical techniques.

Rationality But why is the computed outcome reasonable? Its underlying reasoning involves almost every notion studied in philosophical logic: knowledge about the structure of the game, preferences between outcomes of the game, and also beliefs about what players will do later on in the game, with associated counterfactuals like "if I had played right, ***E*** would have played left". In particular, the key assumption driving ***A***'s reasoning is

Rationality Players only choose actions that they believe to be best for them.

Here the beliefs are about what are the most plausible effects of one's actions in the further course of the game, and 'best' refers to one's preferences among these outcomes.

Observational and theoretical terms The rationality 'formula' entangles action, belief, and preference in an axiomatic manner. I see it as similar to the fundamental axiom $F = ma$ in Newtonian mechanics. The acceleration is observable from the

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press

movements of a physical body, while mass and force are theoretical terms that are only indirectly observable, but when 'imputed' to objects, make for uniform description and ease of calculation. Likewise, we can only observe what people do, but imputing preferences and beliefs creates an explanatory dynamics on top of the pure kinematics of observed actions.[6]

Worries that suggest logical analysis Even so, the results of *BI* may be surprising. For instance, the above outcome *(1, 0)* is not Pareto-optimal: both players end up worse off than in the alternative *(2, 2)*. So, analyzing the precise reasoning underpinning Backward Induction is of interest, and much of the literature looks for assumptions that might be loosened. One well-known result is that *common knowledge of rationality* implies that the *BI* strategy is played (Aumann 1995), but versions with belief also occur. In this setting, we will add a few twists of our own. But in doing so, my aim is not solving any particular puzzle about Backward Induction. I am interested in what styles of reasoning are available about such social settings, and what general issues emerge concerning the logic of interactive agency. We will find a surprising number of these, just by looking at *BI* from various angles.

3 THE LOGICAL FORM OF RATIONALITY: EXPECTATIONS AND THE LOGIC OF BEST ACTION

Relational strategies and set preference Rationality expressed itself in the numerical *BI* algorithm in terms of selecting successor nodes with highest values. But those values arise from a comparison between the *sets* of all possible outcomes of the game compatible with the moves. Now, as is well-known, preferences between sets can be taken in different ways: comparing their minimal values (when one is pessimistic), their maximal values (when one is optimistic), or otherwise. This is our first choice point: there is no unique lifting recipe for set preference that rational players must use. Instead, we find legitimate variety that logics of social action should tolerate. This variety for players will return as we proceed.

While setting set preferences is an interesting degree of freedom, in this paper, we focus on other aspects of game solution. To see these, we will fix one relational version of the *BI* procedure using a strong ∀∀ notion of set preference that goes back to von Wright (1963):

$$X < Y \quad \textit{iff} \ \forall x \in X \, \forall y \in Y : x < y$$

Relational backward induction Here is how Backward Induction works in this relational version. First, mark all moves of the given game tree as active. Call a move *a dominated* if it has a sibling move all of whose reachable endpoints via active

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press

nodes are preferred by the current player to all reachable endpoints via *a* itself. The algorithm then works in stages:

> *At each stage, mark moves as passive whose outcome sets are dominated in the* ∀∀ *sense of set preference by those of another available move, leaving all others active. In this comparison, reachable endpoints are all those that can be reached via a sequence of moves that are still active at this stage.*

The set of moves that are still active at the end form the Backward Induction relation *bi*. Figure 2 is a small illustration, with a variant of an earlier game:

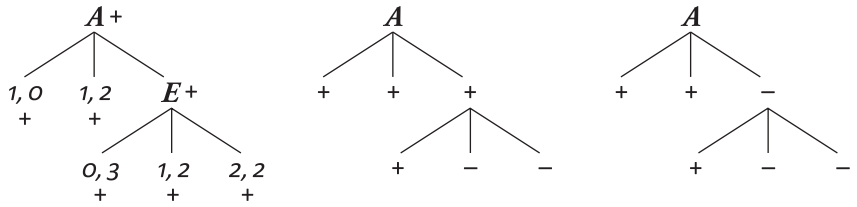


Figure 2

A logical form Many definitions for this relation *bi* exist (cf. van der Hoek & Pauly 2006). We use one from van Benthem, van Otterloo, and Roy (2006):

Theorem On finite extensive games, the *BI* strategy is the largest subrelation σ of the total *move* relation satisfying the following property for all players *i*:

> *RAT* No alternative move for the current player *i* yields outcomes via further play with σ that are all strictly better for *i* than all outcomes resulting from starting at the current move and then playing σ all the way down the tree.

Figure 3 is a useful picture to keep in mind when thinking about what this says:[7]

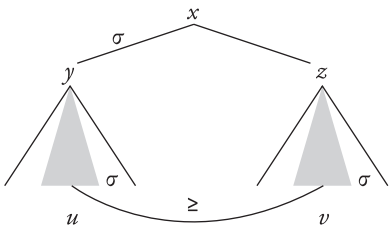


Figure 3

The shaded area is the part that can be reached via further play in our relational strategy.

Reasoning with the modal form of best action? Backward Induction produces, among all my possible moves, my 'best actions'. One natural question is how to capture this level of practical reasoning (one need not always have the details of *BI*

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press

in mind). Consider a modal logic for game trees with modalities $< a >$ for moves as binary relations a on nodes and a preference modality $< pref_i > \varphi$ ("player i prefers some node where φ holds to the current one"). There are some obvious axioms here, stating that *bi* is among the available actions and that preference has its usual properties (cf. Liu 2011). A standard modal semantic technique (Blackburn, de Rijke, and Venema 2000) then yields a more interesting principle:

Fact *RAT* corresponds to validity of the following modal axiom:
$\&_i(turn_i \wedge < \sigma > [\sigma^*](\mathbf{end} \rightarrow p)) \rightarrow [move\text{-}i] < \sigma^* > (\mathbf{end} \wedge < pref_i > p)$

Open Problem What is the complete modal logic of moves, preferences, and best action?

This modal system might be considered the logical core theory of best social action supported by standard rationality. But our very logical form also reveals further difficulties.

Pitfalls of complexity You may think this surface logic is easy to find, but there is a snag. The Rationality in the above picture imposes a structure that entangles two relations: one for actions, another for preference. Such 'grid structure' can make bimodal logics undecidable and non-axiomatizable (cf. van Benthem 2011). There is an interesting tension here. While rationality is an appealing property guaranteeing uniform predictable behavior of agents, it may have a computational cost in high complexity of the resulting logic.[8]

Strategies as beliefs We can also think about the *BI* analysis as yielding histories of a game that players consider *most plausible*. *RAT* then states 'rationality-in-beliefs': *no player plays an action that she believes to be worse than another available one*. There is a precise sense to this equivalence (Baltag, Smets, and Zvesper 2009; van Benthem and Gheerbrant 2010):

Fact There is a one-one correspondence between strategies as subrelations of the *move* relation and connected plausibility orders over all histories of the game.[9]

We will return to this perspective of game solution as a process of creating beliefs later.

We have looked at game solution and found connections with logic. What more could one want? As it happens, we are just at the start of our analysis, and it is time to shift gears.

4. DYNAMIC ANALYSIS: DELIBERATION AS A LOGICAL LIMIT PROCESS

Rational procedure The above analysis still leaves something important unanalyzed. While the relation of 'best action' may be the finished product that Backward Induction delivers, *logical dynamics* (van Benthem 1996, 2011) treats informational procedures as first-class citizens. In particular, expectations in a

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press

game arise from a process of *deliberation*, and the latter itself might be our primary concern. Rationality resides in styles of deliberation and reasoning, not just properties of outcomes. A full-scale introduction to logical dynamics is beyond the scope of this paper. Its 'strong arm' is a family of dynamic-epistemic logics for actions of information update, belief revision, or preference change, with their own logical axioms. In the following scenarios, we will draw on some basic ideas from this approach.

Deliberation as public announcement How to bring the *BI* game procedure into logic? One analysis makes *BI* a process of *prior deliberation* by players whose minds proceed in harmony, though they need not communicate in reality (van Benthem 2007b). Its precise mechanism consists of *public announcements* $!\varphi$ saying that formula φ is true. These transform a current model $(\boldsymbol{M}, s)$ with actual world s into its submodel $(\boldsymbol{M}|\varphi, s)$ whose domain consists of those worlds in $\boldsymbol{M}$ that satisfied φ. The idea for recursive game solution is now to find one assertion whose iterated announcement will prune the game tree until some equilibrium state is reached. For an analogy, think of the way repeated public announcements of ignorance drive the famous epistemic puzzle of the Muddy Children to its resolution.[10]

Local rationality The relevant announcement $!\varphi$ for the deliberative version of *BI* is, not surprisingly, a suitable form of rationality at stages of a game. Recall that, at a turn for player i, move a is *dominated* by move b at the same node if every history through a ends worse, in terms of i's preference, than every history through b. Now φ is

> ***rat*** No player has chosen a strictly dominated move in coming to the present node.[11]

This says that no player has done something stupid. Some nodes satisfy this, others may not. Thus, announcing this formula is informative, and it can make the current tree smaller, leading to new truth values for the formula ***rat*** itself, which can then be announced again, and so on. Consider a game with three turns, four branches, and pay-offs in the order $\boldsymbol{A}$, $\boldsymbol{E}$ (see figure 4).

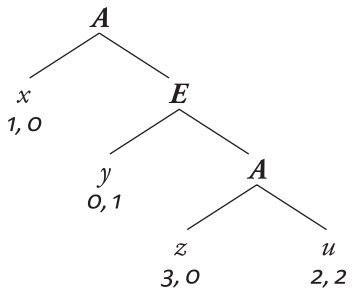


Figure 4

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press

The *BI* path emerges step by step. Stage 0 of the procedure rules out point *u* (the only one where ***rat*** fails), Stage 1 rules out *z* and the node above it (now points where ***rat*** fails), and Stage 2 rules out *y* and the node above it. In the remaining game (shown in figure 5), ***rat*** holds throughout.

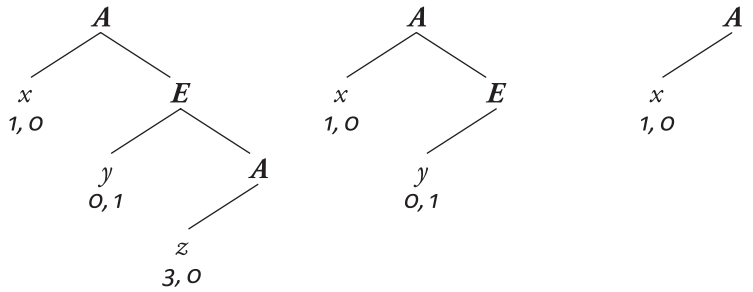


Figure 5

Announcement limits and procedural foundations In this example, a general mechanism is at work. For each model $\boldsymbol{M}$ and formula φ, the *announcement limit* $(\varphi, \boldsymbol{M})^{\#}$ is the first model reached by repeated announcements $!\varphi$ that no longer changes after the next announcement.[12] Either this limit model is non-empty, and φ holds in all nodes, becoming common knowledge (the *self-fulfilling* scenario), or it is empty, and the negation $\neg\varphi$ has become common knowledge (the *self-refuting* scenario). Both kinds occur in actual scenarios: rationality assertions like ***rat*** tend to be self-fulfilling, while the ignorance statement in the Muddy Children puzzle was self-refuting: at the end, it holds nowhere. In this setting, the above iterated public announcements capture Backward Induction:

Theorem In any game tree **M**, the limit submodel $(!\boldsymbol{rat}, \boldsymbol{M})^{\#}$ produced by iterated announcement of rationality ***rat*** is the history computed by the *BI* algorithm.

This limit perspective is not just a source of insight for solving games, but in principle, it applies more broadly to any sort of informational process.[13] Still, open problems abound. We do not know precisely which syntactic formulas are self-fulfilling or self-refuting (on all models), and the study of limit behavior of epistemic assertions is only just starting.

This study is a natural follow-up to epistemic logic, since we usually do not just perform single update steps, but engage in informational processes over time obeying a 'protocol'. This temporal dimension is clear in games, but it also makes sense in epistemology (cf. Hoshi 2009). The dynamic analysis is an alternative to standard foundations of game theory. Repeated announcement of Rationality eventually makes ***rat*** true throughout the limit model: it has *made itself common*

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press

knowledge among all agents. Thus, the latter property is no longer an assumption, it is now *explained* by the dynamic procedure.[14]

5 BELIEFS AND GAME SOLUTION AS ITERATED PLAUSIBILITY UPGRADE

Creating expectations Public announcements are drastic, cutting whole branches of a game tree. Backward Induction might be better construed as creating expectations, rather than knowledge arising from eliminating nodes by 'hard updates' $!\varphi$. Then we need a doxastic 'soft update' that rearranges *plausibility order* between worlds (van Benthem 2007a). In games, this order runs between histories, encoding players' conditional beliefs about what will happen. To do this, we move to another area of logical dynamics: belief change through modification of plausibility orders. A key example is the following operation:

> *Radical upgrade* $\Uparrow\varphi$ makes all φ–worlds best, and puts all $\neg\varphi$–worlds underneath, while keeping the old ordering inside these two zones.

To make this scenario work for *BI*, we adapt the notion of local rationality. Say that move x for player i *dominates* its sibling y *in beliefs* if the *most plausible* end nodes reachable after x along any path in the whole game tree are all better for i than all the *most plausible* end nodes reachable in the game after y. *Rationality-in-beliefs* is then the assertion

> ***rat**** No player plays a move that is dominated in beliefs.

Now we perform a relation change that is like a radical upgrade $\Uparrow$***rat****:[15]

> If x dominates y in beliefs, we make all end nodes from x more plausible than those reachable from y, keeping the old order inside these zones.

This changes the plausibility order, and hence the dominance pattern, so iteration makes sense. Figure 6 shows the stages in an earlier game, where x, y, z stand for end nodes or histories:

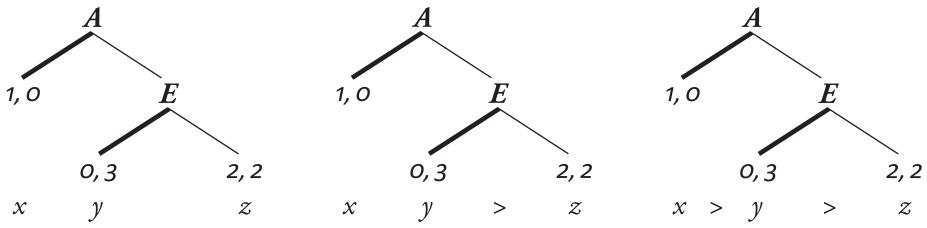


Figure 6

In the first tree, going right is not yet dominated in beliefs for ***A*** by going left, ***rat**** only affects ***E***'s turn, and the update makes *(0, 3)* more plausible than *(2, 2)*. But then, going right is dominated in beliefs, and the next update makes ***A*** going left

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press

most plausible.[16] Now, in terms of an earlier correspondence between strategies and beliefs, we have:

Theorem On finite trees, the Backward Induction strategy is encoded precisely in the plausibility order for end nodes (i.e., histories) created in the limit model $(\Uparrow \boldsymbol{rat}^*, \boldsymbol{M})^{\#}$ of iterated radical upgrades with rationality-in-belief.

Again the usual characterization assumptions of game theory fall out for free. At the end of this procedure, players have acquired *common belief in rationality*.

The road ahead: belief revision and learning theory Once more, something general is at work here: the limit behavior of upgrades changing plausibility patterns. This may go beyond the self-fulfilling and self-refuting cases for repeated public announcements. Baltag and Smets (2009) show how some repeated upgrades $\Uparrow\varphi$ lead to *cycles* of plausibility – and we have no taxonomy of such tricky statements φ. The limit behavior of creating expectations seems a new area of investigation.[17] An interesting link here is with *formal learning theory* (cf. Kelly 1996). Belief revision procedures are learning functions that produce varying hypotheses over time. Uniform learning methods then match up with our iterated upgrades, and it can even be shown that radical upgrade is a universal format.[18]

6 FROM DELIBERATION TO ACTUAL PLAY: THE INVERSION PROBLEM

Our discussion so far may have suggested that Backward Induction is the view of games officially endorsed by logic. But the opposite is true. The above limit scenarios $(!\varphi)^{\#}$ and $(\Uparrow\varphi)^{\#}$ work with any formula φ in our logical language, whether it produces the *BI* solution or not. And there are even further, more radical ways of bringing logic into play here:

From prior deliberation to posterior update Backward Induction is just one way of creating plausibility in a game. Its limitations show vividly in the following scenario, which one might dub the 'inversion problem'. So far, we viewed *BI* as a deliberation procedure to be followed prior to playing the actual game. But what about the real a posteriori dynamics of playing a game, where we observe moves (hard information in the above sense) while also changing our expectations (perhaps through soft updates in the above sense)?

One might think this is a simple matter of inversion, following the 'virtual moves' computed in the deliberation phase. But often *BI* makes little sense when run in this opposite direction. Assuming the deliberation phase, we expect a player to follow the *BI* path. So, if she does not, we must revise our beliefs – and one way of doing that is precisely having second thoughts about her style of deliberative reasoning.

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press

The paradox of backward induction Why would a player who has deviated from *BI* return to *BI* later on (Bicchieri 1988)? This may be very unlikely. Consider the example depicted in figure 7:

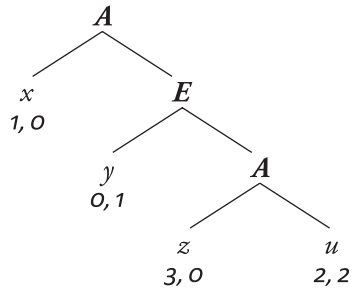


Figure 7

Backward Induction tells us that ***A*** will go left at the start. So, if ***A*** plays *right*, what should ***E*** conclude? One might still assume that he will play *BI* afterwards, thinking that ***A*** made a single mistake. Or one can be optimistic, with an unshakable faith in future rationality. These assumptions are the technical core of the results by Aumann (1995) and similar work within dynamic-epistemic logic by Baltag, Smets, and Zvesper (2009), that characterize the *bi* strategy.

Many belief revision policies But clearly, this is not the only way of revising one's beliefs here. ***E*** could have other reasonable responses, like

> "***A*** is trying to tell me that he wants me to go right, and I will be rewarded",
> "***A*** is an automaton with a rightward tendency, and cannot act otherwise",
> "***A*** believes that ***E*** will go right all the time",

and so on. Our dynamic logics do not choose for the agent. They support many belief revision policies, of which *BI* is just one (Stalnaker 1999). Our earlier analyses did not yet highlight this degree of freedom for players, since the *BI* algorithm made uniform assumptions about players' prior deliberation and what they will do as the game proceeds.

The inversion problem is a central challenge to a logic of social action. Our expectations may be based on prior deliberation, including scenarios that we think will not occur, but what if the unexpected happens? A move considered hypothetically may impact us quite differently once it has occurred. Thus, can a priori styles of deliberation be in harmony with actual play of a game that includes updating in response to what happens? There are often deviations from this harmony in practice,[19] but it is definitely an interesting case to understand better.

In general, as we noted, real-time play of a game may involve very different hypotheses about the kind of player one is dealing with. In what follows, we discuss one particular line, pursuing styles of analysis inspired by game theory that stay close to rationality.

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press

7 UPDATING BELIEFS FROM OBSERVATIONS

Factoring in the past Let us try to change Backward Induction to make it more robust against inversion. Here is the obvious idea. In earlier computations, we omitted the past history: it did not matter for our future expectations how we arrived at the present node. But if we think that the past can be informative, then we should factor it in.

Rationalizing a game Figure 8 is a simple example, a variant of an earlier game:

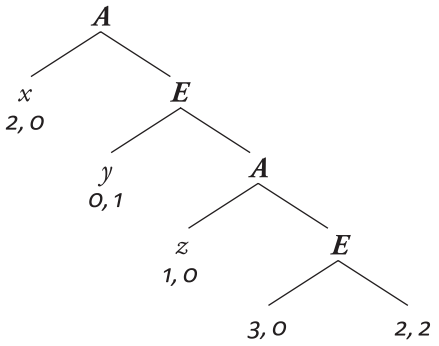


Figure 8

> Suppose ***A*** moves *right* at the start. Assuming that ***A*** is rational-in-beliefs, this tells ***E*** something about his beliefs and/or intentions. Clearly, ***A*** does not expect ***E*** to go left at her first turn, because then playing *left* at the start would have been better. For the same reason, ***A*** does not intend to go *left* after that at his second turn. And we may assume he believes that ***E*** will go *right* at the end, as that is the only way his opening move makes sense. This might induce ***E*** to go right at her first move – though one hesitates to predict what she will do at the end.

The point here is not that we have one simple rule replacing *BI*. It is rather that *the past is informative*, telling us which choices players made or avoided in coming here.[20]

'Games with a history' The main change needed looks simple, but it is not. We now look at games ***M*** with a distinguished point *s* indicating how far the actual play has progressed (figure 9).

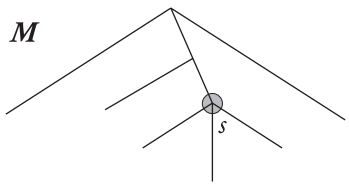


Figure 9

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press

Thus, at the actual stage, we know what has happened, and players can let their behavior depend on a mixture of two things: what the remaining game looks like, and what players have done so far in the context of the larger game. In particular, this means that, unlike with Backward Induction, which created one uniform binary plausibility relation $x \leq y$ among histories x, y, we now get a *ternary plausibility relation* $\leq_s xy$ (van Benthem 2004).

Ways of taking observed behavior In terms of this picture, how will players change their expectations as the black dot moves along the game? There is no unique prescribed recipe. Figure 10 presents a few simple scenarios.

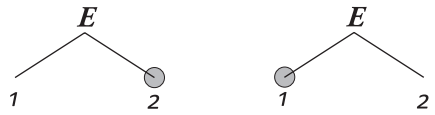


Figure 10

In the first scenario, we see the end of a basic rational decision. In the second, we see a 'stupid move', probably regretted by ***E*** once made.[21] Figure 11 illustrates a few more complex games with a history.

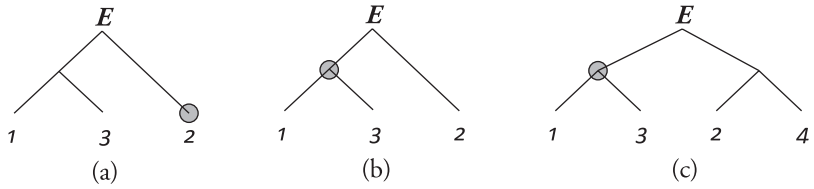


Figure 11

The play observed in Game (a) may be considered rational by ascribing a belief to ***E*** that choosing left would have resulted in outcome 1. Game (b) may also be rationalized by ascribing a belief to ***E*** that the game will now reach 3. Finally, Game (c) suggests that ***E*** thinks she will reach 3, while she would have reached 2 if she had gone right.

Some choice points There are many options for making sense of the observed behavior. Actual moves may be considered stupid. They may also be taken to be smart – but how smart again depends on assumptions about players. Here is a natural stipulation:

> ***Rationalize*** By playing a move, a player gives information about her beliefs. These beliefs are such as to rule out that her actual move is strictly dominated-in-beliefs.

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press

This will only work if the player does not choose a move that is strictly dominated under all circumstances, i.e., under all possible continuations of the game. This 'weak rationality' avoids stupid things. One could also assume 'strong rationality', where the agent thinks her current move is best for her. Next, effects of such update rules depend on assumptions about the belief structure of the agent. With minimal rationality and beliefs allowing for ties, one way of rationalizing is to assume that the agent considers all continuations equally plausible. The move can then never be strictly dominated in beliefs. But assuming that agents always have one unique most plausible history in mind, more information may come out of an observed move.[22] By fixing such stipulations, we get various rationalization algorithms, all proceeding on the principle that moves reveal beliefs about the future.[23]

Discussion Revealed 'beliefs' of a player ***E*** do double duty. Connected to a turn for the other player ***A***, they correspond to real beliefs about what ***A*** will do. But with a turn for ***E*** herself, they are like *intentions*. Next, the ternary plausibility order allows for differences between what players would expect hypothetically if another move had been played than the actual one ('that would have been a stupid act') versus how they would feel if that other move were actually played. Backward Induction had no such distinction – but with general rationalization, the algorithm need not produces expectations that 'match' across the game.[24] There need not even be 'monotonicity' going down an observed history: expectations may still change, e.g., when a player makes an irrevocably stupid final move.[25]

Some general issues To bring all this into a dynamic logic, several things must happen. One is to define precise updates for rationalization. It is not clear that these work just on game trees. There may be quite different belief structures for an agent that rationalizes a given move, which need not have a 'weakest common case'. Hence, we may have to move to more complex models consisting of game trees with different belief structures.[26] There is also the issue of *iteration*: rules for successive belief changes along a path – but we stop here. Our final point is triggered by an obvious practical question:

Can dropping BI be for the best? In extensive games, rationalizing hits difficulties. At final stages, we may expect the last player to be rational, whatever the course of the game has taught her, and then it is hard to avoid running into the *BI* moves after all. That is why some authors on Forward Induction look at extensive games that end in strategic matrix games, or more simply, that have final stages with *simultaneous moves*. Figure 12 contains a concrete example where dropping *BI* then becomes advisable, adapted from (Perea 2011):

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press

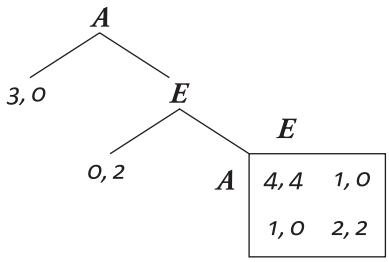


Figure 12

In the matrix game to the right, no move for any player dominates any other move. So ***E*** considers all outcomes possible if we apply Backward Induction reasoning. But then going left is safer for her than going right, and hence ***A*** should go left at the start. However, if we rationalize, and see ***A*** go right, then ***E*** will have extra information at her choice point: ***A*** expects to do better than the outcome 3, which can only be because he expects to play *Up* in the matrix game. Now this tells ***E*** that she should go there and play *Left*, doing better than her earlier *BI*-move giving her just 2.

While this story is plausible, it needs a convincing interpretation for the final matrix game. Here is one. Often we do not know the complete game, or it is beyond our powers to represent. We only know some initial structure. The matrix game of imperfect information then gives a rough image of what might happen afterwards.[27]

We started by analyzing Backward Induction as a process of deliberation prior to playing a game. We have now discussed how to analyze a game as it is being played. At the other extreme would be a process of rationalizing a game that has already been played to its end.

8 DIGRESSION: RATIONALIZING BY UPDATING PREFERENCES

So far we assumed that players have at least minimal rationality, avoiding totally dominated moves. But we can even make sense of irrational behavior.

Updating preferences Our next ploy is viewing observed moves as revealing information about players' preferences. If we just observe actions of a player, constructing preferences on the fly, any behavior can be rationalized. Here is one of many folklore results:

Theorem Any strategy against the strategy of another player with known preferences can be rationalized by assigning suitable preferences among outcomes.

Here is how to do this, working inductively bottom-up:

Let ***E*** be the player whose behavior is to be rationalized. Assume that all subgames for moves have been rationalized already. Now consider the actual move *a* made by ***E*** and its subgame G_a. We make ***E*** prefer its outcome more than that of any subgame G_b for

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press

another available move *b*. To do so, we just add a number N to all values assigned already to outcomes in G_a. By making N large enough, we can get any outcome in G_a to majorize all outcomes in subgame G_b. Here adding the same number to all outcome values in G_a does not change any earlier relative preference in that subgame. Moving upward to turns for the other player ***A***, there is nothing that we need to adjust for ***E***.

If we also assume that the player ***A*** whose preferences are given is 'minimally rational', never choosing a move strictly dominated in all its possible outcomes by another available one, we can even assign preferences to ***A*** to match up with Backward Induction play.

Dynamic logics of preference change The background to the above algorithm are dynamic logics of *preference change*. The updates in our algorithm give information about some player's preferences, not setting those preferences themselves – but the two views are close. Rationalization may be seen as successive preference changes following observations of moves of a game.[28,29]

We are on new logical territory here: the first monograph on dynamic logics of preference change seems to be (Liu 2011) (but cf. also Girard 2008).

Coda: changing the game itself On this road to realism there are many more options. Players may not know the game that they are playing, a common scenario in social life. And even if they do know the game, they may want to change it. Van Benthem (2007c) discusses *promises* changing games, and gives dynamic-epistemic logics for these.[30] Other game changes affect players' options by adding or deleting moves, affecting 'best action'. We may even remove or add players, changing individual powers and possible coalitions. The matching dynamic logic is a neglected topic: e.g., we know little about stability of strategies across changing games. But such comparisons seem crucial to our enterprise. Recall the earlier analogy with mechanics. Why is postulating a force function on top of observed behavior more than an ad-hoc device for making Newton's axioms true? The point is that this move still yields good results when we change the physical situation, adding or removing objects. Getting to grips with such uniformities is a major challenge to the logical analysis of games.

9 LOGIC+GAME THEORY = THEORY OF PLAY

Games plus players on a par We have seen a diversity of construing behavior in games. This seems right, since in social action we do all these things: predicting the future, explaining past behavior, and so on. But this also means there is no unique principled way of defining 'best action'. There is a missing ingredient, and it is a typical social feature: information about the types of agent we are interacting with. The structure of a game by itself does not provide this information, unless we make strong uniformity assumptions. We need more input. The term coined for this in (van Benthem, Pacuit, and Roy 2011) is '*Theory of Play*', an extension of 'Game Theory'. To make sense of what happens in a game, we must combine information

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press

about game structure with information about the agents in play. Game theory allows each player her own preferences, but it imposes uniformity on how players think and act; witness the symmetries in the Backward Induction algorithm.[31] But we need much more variety: in computational limitations, belief revision policies, etc.[32]

At present, there is no Theory of Play: only bits and pieces of logical and game-theoretic equipment that might help us create one. There can be different observations by players of imperfect information games, different belief revisions following observations, and different changes in preference. Theory of Play should also allow for different hypotheses about the capacities of players. The dynamic logics that we have used in the above do tolerate such diversity, and sometimes help identify possible points of variation.[33]

Discussion: complexity and messiness Theory of Play is rich, but it comes at a cost. It requires maintaining a much larger space of hypotheses about agents, and hence of models that can be much more complex than the simple game pictures that we have drawn so far.[34] There is also the worry of 'messiness': logical systems that acknowledge variety tend to get complex.[35] But this may be a matter of choosing the right architecture. An example is the dynamic logic of belief revision. Treated prima facie (van Benthem 2007a), it dissolves into a jungle of different sorts of relational update, with complete logics for each kind. But at a higher abstraction level (Baltag and Smets 2008), there is one rule of Priority Update over richer 'event models' encoding the variety, with just one logic that can be axiomatized and understood in a simple manner. The challenge of Theory of Play is acknowledging the diversity, while still letting logic do its usual job of abstraction and idealization.

10 REPERCUSSIONS FOR LOGIC ITSELF

I conclude with a few thoughts on what the above means for logic itself. First, of course, the positive take on 'messiness' is the attraction of a rich area of investigation. Moreover, the themes we have seen give a nice 'boost' to traditional topics in philosophical logic.

Must it all be dynamic? The emphasis in this paper has been on dynamics: a Theory of Play as a model for social action imports individual and social actions into logical systems. But this is just to forcefully put a new perspective on the map. The dynamic reality of intelligent social behavior contains two crucial directions: *zooming in* on greater detail, *and zooming out* to broader views of a situation. From the latter point of view, traditional static logics of belief, or our 'best action', still represent entirely viable and useful concerns.

Taking Theory of Play inside logic Theory of Play is not just a concern about games or social action. Taken seriously, it also acknowledges agent diversity inside

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press

the heartland of logic, viz. the study of reasoning itself. Proof Theory is a study of idealized proofs by idealized agents. What about a 'Theory of Inference' describing human or computational agents that engage in deduction and other activities, and their different styles of doing so? Is this just messy reality close to cognitive science, or could logic have more to say about the mechanics of actual reasoning if we bring to light the uniformity assumptions in current systems like propositional and predicate logic, and then try to relax them?[36]

All this suggests a perhaps even more radical perspective. While this paper is about logic *of social action*, logic does not stand apart from this field of study. Reasoning itself is also a typical social action, from discussions between philosophers to research groups in science. What are the repercussions of our present stance for logic *as social action*? This would go well beyond the boundary of this paper, but I think this road, too, is worth exploring.

REFERENCES

Aumann, R. 1995. "Backward Induction and Common Knowledge of Rationality." *Games and Economic Behavior* 8(1): 6–19.

Baltag, A. and S. Smets. 2008. "A Qualitative Theory of Dynamic Interactive Belief Revision." In G. Bonanno, W. van der Hoek, and M. Wooldridge (eds.), *Texts in Logic and Games Vol. 3*, pp. 9–58. Amsterdam: Amsterdam University Press.

Baltag, A. and S. Smets. 2009. "Group Belief Dynamics under Iterated Revision: Fixed Points and Cycles of Joint Upgrades." *Proceedings of the 12th Conference on Theoretical Aspects of Rationality and Knowledge*, pp. 41–50. New York: ACM.

Baltag, A., S. Smets, and J. Zvesper. 2009. "Keep 'Hoping' for Rationality: A Solution to the Backward Induction Paradox." *Synthese* 169: 301–33.

Bicchieri, C. 1988. "Common Knowledge and Backward Induction: A Solution to the Paradox." *Proceedings of the 2nd Conference on Theoretical Aspects of Reasoning about Knowledge*, pp. 381–93. Los Altos, CA: Morgan Kaufmann Publishers.

Blackburn, P., M. de Rijke, and Y. Venema. 2000. *Modal Logic*. Cambridge: Cambridge University Press.

Brandenburger, A. 2007. "Forward Induction." Manuscript, Stern School of Business, New York University.

Dégrémont, C., L. Kurzen, and J. Szymanik. 2011. "Cognitive Plausibility of Epistemic Models: Exploring Tractability Borders in Epistemic Tasks." Manuscript, ILLC, University of Amsterdam.

Dégrémont, C. and O. Roy. 2009. "Agreement Theorems in Dynamic Epistemic Logic." In A. Heifetz (ed.), *Proceedings of the 12th Conference on Theoretical Aspects of Rationality and Knowledge*, pp. 91–8. New York: ACM.

Gheerbrant, A. 2010. *Fixed-Point Logics on Trees*. Dissertation, ILLC, University of Amsterdam.

Gierasimczuk, N. 2010. *Knowing One's Limits: Logical Analysis of Inductive Inference*. Dissertation, ILLC, University of Amsterdam.

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press

Girard, P. 2008. *Modal Logic for Belief and Preference Change*. Dissertation, Department of Philosophy, Stanford University, and ILLC, University of Amsterdam.

Halpern, J. and M. Vardi. 1989. "The Complexity of Reasoning about Knowledge and Time, I: Lower Bounds." *Journal of Computer and System Sciences* 38: 195–237.

Hoshi, T. 2009. *Epistemic Dynamics and Protocol Information*. Dissertation, Department of Philosophy, Stanford University.

Joyce, J. 2004. "Bayesianism." In A. Mele and P. Rawling (eds.), *The Oxford Handbook of Rationality*, pp. 132–55. Oxford: Oxford University Press.

Kelly, K. 1996. *The Logic of Reliable Inquiry*. Oxford: Oxford University Press.

Liu, F. 2011. *Reasoning about Preference Dynamics*. Dordrecht: Springer.

Parikh, R., C. Tasdemir, and A. Witzel 2011. "The Power of Knowledge in Games." Working paper, CUNY Graduate Center and New York University.

Perea, A. 2011. "Belief in the Opponents' Future Rationality." Working paper, Epicenter, Department of Quantitative Economics, University of Maastricht.

Stalnaker, R. 1999. "Extensive and Strategic Form: Games and Models for Games." *Research in Economics* 53: 293–319.

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

van Benthem, J. 2004. "Update and Revision in Games." Lecture notes, ILLC, University of Amsterdam and Stanford University.

van Benthem, J. 2007a. "Dynamic Logic of Belief Revision." *Journal of Applied Non-Classical Logics* 17: 129–55.

van Benthem, J. 2007b. "Rational Dynamics." *International Game Theory Review* 9(1): 13–45. Erratum reprint, 9(2): 377–409.

van Benthem, J. 2007c. "*Rationalizations and Promises in Games*." *Philosophical Trends*, Supplement 2006, pp. 1–6. Beijing: Chinese Academy of Social Sciences.

van Benthem, J. 2011. *Logical Dynamics of Information and Interaction*. Cambridge: Cambridge University Press.

van Benthem, J., J. Gerbrandy, T. Hoshi, and E. Pacuit. 2009. "Merging Frameworks for Interaction." *Journal of Philosophical Logic* 38: 491–526.

van Benthem, J. and A. Gheerbrant. 2010. "Game Solution, Epistemic Dynamics and Fixed-Point Logics." *Fundamenta Informaticae* 100: 1–23.

van Benthem, J., S. van Otterloo, and O. Roy. 2006. "Preference Logic, Conditionals, and Solution Concepts in Games." In H. Lagerlund, S. Lindström, and R. Sliwinski (eds.), *Modality Matters*, pp. 61–76. Uppsala: University of Uppsala.

van Benthem, J., E. Pacuit, and O. Roy. 2011. "Games and Interaction: The Logical Perspective." *Games* 2(1): 52–86.

van der Hoek, W. and M. Pauly. 2006. "Modal Logic for Games and Information." In P. Blackburn, J. van Benthem, and F. Wolter (eds.), *Handbook of Modal Logic*, pp. 1077–148. Amsterdam: Elsevier.

von Wright, G. H. 1963. *The Logic of Preference*. Edinburgh: Edinburgh University Press.

NOTES

1 I would like to thank Fenrong Liu, Eric Pacuit, and Paul Weirich for their helpful comments.

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press

2 This paper will not be concerned with probabilistic aspects, though these might be added in later.
3 I will ignore the important issue of 'first-person' vs. 'third-person' perspectives henceforth.
4 This paper will be largely about finite games of perfect information, though many of our results would also apply to games with imperfect information.
5 In the tree, the moves of the relevant strategy are marked by bold-face lines.
6 Matters of retrieving beliefs and preferences from behavior have been studied much more deeply in the philosophical foundations of statistics and decision theory; cf. (Joyce 2004).
7 The property depicted here may be stated formally as $\&_i\ \forall x\ \forall y\ ((Turn_i(x) \wedge x\sigma y) \rightarrow (x\ move\ y \wedge \forall z\ (x\ move\ z \rightarrow \exists u\ \exists v\ (end(u) \wedge end(v) \wedge y\ \sigma^*\ v \wedge z^*\sigma^* u \wedge u \leq_i v)))$. Gheerbrant (2010) has an extensive analysis of this and other variants of relational *BI* in fixed-point logics of computation.
8 A well-known case involves agents with *Perfect Memory*. These show a regular geometric interplay of knowledge and action, but the mathematics of regular structures can be rich, and so, their epistemic action logic is of high non-axiomatizable complexity (Halpern and Vardi 1989). However, a recent concern has been the distance between the complexity of such a theory for reasoning about agents and the complexity of tasks for the agents themselves; cf. (Dégrémont, Kurzen, and Szymanik 2011).
9 A small technical condition of 'node compatibility' on the plausibility orders is omitted here.
10 A father tells his children that at least one of them is dirty. He then asks if they know their status. After initial rounds of ignorance, the muddy children learn their status: after that, the clean ones do.
11 This assertion can be formulated explicitly in extended hybrid modal languages of game trees.
12 Such a limit exists in finite models since the sequence of submodels is weakly decreasing. Announcement limits also exist in infinite models, by taking intersections at ordinal limit stages.
13 Dégrémont and Roy (2009) analyze repeated public communication of *disagreement* between agents, and show how this is often self-refuting: in the limit, agreement is achieved.
14 Admittedly, the dynamic procedure also builds in rationality, in its local steps.
15 We omit some technical details here: plausibility upgrades must be 'local' inside subgames.
16 We can think of this shared order as players' *intentions* for their own actions and their *expectations* about what others will do. Perhaps surprisingly, *BI* makes all these uniform.
17 We need perspicuous logics that define limit submodels obtained by these procedures.
18 (Gierasimczuk 2010) has technical details and more extensive motivations.
19 There is an aspect of 'cold feet' here. It is easy to be brave when reasoning about future events, but it is hard to stick to one's plan when the unexpected has become real. One can see this as a struggle between hypothetical and temporal 'if', questioning the usual equivalences (even in dynamic-epistemic logic!) between conditional beliefs and beliefs after some fact has materialized.

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press

20 Scenarios like this often go under the name of 'Forward Induction' (cf. Brandenburger 2007).
21 Not every played move must be best. Being able to do things we regret is a key feature of games.
22 Unique belief plus strong rationality were in fact the reasons for suggesting that in the above Game (c), the agent believes that she will get 3, and would have got only 2 when going right.
23 We will not pursue such algorithms in this paper, but (van Benthem 2007c) has examples.
24 In our current procedure, we only use off-path expectations in the game as a contrast allowing us to get more information about relevant beliefs in the future of our current path.
25 Backward Induction might be the *only* algorithm that creates uniform and monotonic expectations.
26 This need for 'lifting' is known for reasoning about strategies, cf. (van Benthem 2011, ch. 10).
27 A game logic for hybrids of extensive and strategic games would allow simultaneous moves.
28 Upgrades are a bit more complex than in the above procedure: see (Liu 2011) for details.
29 Adjusting preferences also works if the *beliefs* of the player are given beforehand, since as we have seen earlier, a relational strategy or a belief amount to the same thing.
30 Parikh, Tasdemir, and Witzel (2011) discuss agents manipulating knowledge during play.
31 Game Theory does distinguish 'competitive' and 'cooperative' games. But most social situations are a bit of both, depending on the players, and we must learn about the mixture as we proceed.
32 The uniform case remains important. Some ethical theories find it a moral duty to treat agents as 'equal'. A Theory of Play takes no stand, but would ask this: Do we take other agents seriously if we treat them as being as well-informed and logically endowed as we are? Or is morality the art of being reasonable to others in a diverse society? An answer to these issues may depend on the game. Should 'fair' exams be adapted to the intelligence level of individual students? I do not think so.
33 One such identification is the representation theorem in (van Benthem et al. 2009) for dynamic-epistemic temporal update over time induced by standard 'product update'. This turns out to be precisely the historical record of informational behavior for agents endowed with Perfect Memory, and a No Miracles property of learning from new observations only. But one can also have update logics for highly memory-bounded agents such as finite automata (cf. Liu 2011).
34 Van Benthem (2011) proposes a hierarchy of three levels of 'worlds' for logics of social action: nodes in game trees, histories in game trees, and thicker possible worlds that encode games, strategy profiles, and other relevant features. Theory of Play seems to need all three.
35 *A worry here.* Models of social action in a Theory of Play require more and more sophistication. But does not daily life show that we need *less*, rather than *more* of this? Here is another avenue of exploration when facing the Inversion Problem. What about a simple behaviorist stance? You have done certain things. I am not going to delve deeply into hypotheses as to why, but I stick to some simple rules of commitment. If

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press

you have done something good for me, you achieved an *entitlement*, and my score of the game will tell me how to respond. Entitlement versions of game solution are sketched in (van Benthem 2007b), drawing an analogy between Backward Induction and 'forward-tinkering' views of society like that of Rawls, and between commitment procedures and historical entitlement theories in the style of Nozick. But there is no conflict: I think we can pursue both.

36 Such a theory might require one even further step. Should we also *explicitly represent agents*, say in the form of automata? Then we might parametrize the monolith of first-order reasoning: as performed by finite automata, by push-down-store automata, or, in the limit, by Turing machines.

Johan van Benthem is a University Professor of Logic at the University of Amsterdam, Henry Waldgrave Stuart Professor of Philosophy at Stanford University, and Distinguished National Expert in the Humanities at Tsinghua University. His current main fields of interest are logic, information, and interaction, including connections with philosophy, computer science, game theory, and cognitive science.

https://doi.org/10.3366/epi.2011.0019 Published online by Cambridge University Press