An abstract approach to reasoning about games with mistaken and changing beliefs

Benedikt Löwe[*]

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Plantage Muidergracht 24, 1018 TV Amsterdam, The
Netherlands

Department Mathematik
Universität Hamburg
Bundesstrasse 55, 20146 Hamburg, Germany

Mathematisches Institut
Rheinische Friedrich-Wilhelms-Universität Bonn
Beringstrasse 1, 53115 Bonn, Germany

bloewe@science.uva.nl

Eric Pacuit

Department of Philosophy
Stanford University
Stanford CA 94305-2155, United States of America

epacuit@stanford.edu

1 INTRODUCTION

Formalizing and understanding reasoning processes in multi–agent situations with imperfect information is one of the foremost tasks of modern artificial intelligence research. After its early phase ("good old–fashioned AI" or "GOFAI"

---

using a phrase of John Haugeland), artificial intelligence did not pay much attention to logic as a means towards achieving this end; logic and artificial intelligence grew increasingly apart. In the past years, this development has reversed: logic overcame its traditional focus and started to study phenomena of interaction and interactive reasoning. This is best represented in the research area called *Logic and Games* in which logicians used dynamic techniques in order to study behaviour in game and communication situations and in what Parikh calls *Social Software* [Par02]. These new developments allowed logicians to provide theoretical insight in the general project of understanding reasoning processes in multi–agent situations. As representative examples from the vast literature, we should like to mention van Ditmarsch's complete epistemic formalization of the game Cluedo® [vD00] and the Broersen–Dignum–Dignum–Meyer deontic logic of deadlines [BDDM04].

These development have triggered a fruitful interaction between the *Logic and Games* community and game and decision theorists who always had an interest in analyzing what constitutes rational behaviour under uncertainty. One particular interesting encounter between logic and game theory is the use of belief revision techniques in the sense of [Gär92] as a means of analysis of games.[1] The game–theoretic analysis of rationality and the study of belief revision have in common that they have a normative hue; they are concerned with questions of what constitutes rational behaviour and what would be quality measures for rationality.

In this paper, we take a decidedly different route. We do not believe that logic is the sole answer to deep and intriguing questions about human behaviour, but we think that it might be a useful tool in simulating and understanding it to a certain degree and in specifically restricted areas of application. We do not aim to resolve the question of what rational behaviour in games with mistaken and changing beliefs is. Rather, we develop a formal and abstract framework that allows us to reason about behaviour in games with mistaken and changing beliefs leaving aside normative questions concerning whether the agents are behaving "rationally"; we focus on what the agents *do* in a game. In this paper, we are not concerned with the reasoning process of the (ideal) economic agent; rather, our intended application is *artificial* agents, e.g., autonomous agents interacting with a human user or with each other as part of a computer game or in a virtual world. Arguably, when such agents interact, the underlying epistemic and rationality assumptions are much less important than the *actual* reasoning process used by the agents. Thus we as-

---

[1]The following is a list of relevant papers combining epistemic logic and game theory: [Aum99], [BS03], [BB99], [Bra07], [dB04], [MBGVS97], [Pea84], [Sta98]. More specifically, the papers [Sta98], [Boa04], and [Per] have pointed out the importance of belief revision in the context of reasoning about solution concepts in games.

sume the players are "programmed" to execute Zermelo's backward induction algorithm when it is their turn to move.[2]

We assume that each player has his or her own set of beliefs about the preferences of the other players. It is with respect to these subjective preferences that the players execute the backward induction algorithm assuming that the preferences and beliefs of the players change over time. Our model highlights what happens when such agents interact in the presence of changing beliefs and preferences. Questions about rationality and the choice of belief revision functions are left open to be determined in tune with the concrete application by those who wish to apply our framework.

In § 2, we give a story of mistaken beliefs that is a typical example of the situation in which we should want our formal setting to be applied. Then, in § 3, we give the definitions for our formal system and how to use this setting to get a backward induction solution. In the following § 4, we then apply our semantics to the story related in § 2 and give an analysis of it.

Our § 5 contains a discussion of related work and future projects. We discuss the advantages of our approach over existing approaches and indicate how it can be connected to the existing literature.

2   A STORY OF REASONING WITH FALSE BELIEFS

The following is a fictitious story in the style of a TV drama. The reader can imagine that this is the outline of a script. The reasoning processes referred to in the story can be made visible to the audience by monologues (Walter talking to himself in his car) or by conversations with some *confidant* or *confidante*.

> *Sue and Jeff have known each other for years. They studied computer science together in the 1980s, and both started their own software companies in the 1990s. Sue is married to Walter, an artist, and Jeff is married to Mary. In the past years, Sue and Mary have become best friends.*
>
> *However, unbeknownst to Sue, her husband Walter and Mary have an affair. Walter, being absolutely dependant on the money of his wife, has no intention of leaving her at all, and wants to avoid that she gets to know*

---

[2]Backwards induction is the most intuitive method for solving extensive games. It was first employed by Zermelo [Zer13] to show that the game of chess is determined and has since then been extended to infinite games [GS53] and found many applications in game theory, mathematics and computer science. Despite its naturalness, the rationality of backward induction has been questioned, for instance in Rosenthal's centipede games [Ros81], and in fact, experimental results show that human players do not follow the backward induction strategy in these games [MP92, NT98]. Stemming from an interesting debate between Ken Binmore [Bin96, Bin97] and Robert Aumann [Aum95] about the underlying decision–theoretic assumptions of the backward induction algorithm, much of the literature has focused on the epistemic conditions (*e.g.*, *common knowledge*) and rationality assumptions (*e.g.*, the players maximize subjective expected utility) that *guarantee* the players follow the backward induction solution. *Cf.* [Per07] for a survey of the relevant literature and § 5.2 for some ideas how to reconcile our framework with this line of research.

about this affair at all costs. He believes that the fact that Sue and Mary are best friends acts as a safeguard for his secret: Mary must be aware that she will lose Sue as a friend if Sue finds out, and Mary clearly doesn't want that. So, Walter convinced himself that Mary will never ask him to separate from Sue or—even worse—tell Sue about the affair.

Mary on the other hand is rather unhappy with Jeff, and really wants to leave him. She believes that her friendship with Sue is robust enough to survive the fact that she has an affair with Walter. In her dreams, she imagines a nice future with Walter. She is convinced that if she presses Walter enough, he will finally leave Sue for her. She can make up with Sue afterwards.

One morning, she gives Walter an ultimatum: he should make up his mind and choose between her and Sue. Walter is ultimately confused: he must have misjudged Mary. Stuck in the traffic jam on his way to an appointment with a potential client, his mind raced: If he chooses Mary, then Sue would know about their affair, and Mary would lose her best friend. What was Mary thinking? The only rational explanation that he could come up with was that Mary wanted to be with him so badly that she would give up her friendship with Sue for it and expected him to say 'yes' to her ultimatum… Obviously, for financial reasons, Walter couldn't leave Sue. But he needed to be careful here: if he said no to Mary, would she tell Sue? No, he reasoned, since then she would lose both Sue and him which is definitely worse than just losing him. So, he'd be safe. Smiling, he used his cell phone to call Mary and tell her that he would not leave Sue.

When she hung up the phone, Mary was fuming with anger. Apparently, Walter wanted to stay with Sue. "Well, if that's what he wants, then I must have completely misjudged him. I should cut my losses, and at least be honest to my best friend and tell her," she reasoned, and acted accordingly.

And Mary was right in her judgement of Sue. The two women discussed the matter, and when Walter returned from his appointment in the afternoon, his paintings were standing on the front lawn of their house and the lock of the front door had been changed. Walter gazed emptily at his paintings searching for a logician to help him to figure out what had happened.

We should stress that human beings have no problems in analysing an episode like this—with ease, they can make judgements like "Walter is wrong about his judgement of Sue and Mary; after the ultimatum, there was no chance of staying together with Sue anymore, but he could have saved his relationship with Mary, hadn't he misjudged his wife".

For computational situations, we should like to be able to do the same within some preferably simple formal system. Being able to formally access the reasoning structure of episodes like this is crucial for the analysis of games with mistaken beliefs.

3 FORMALIZATION

Let I be the finite set of players whom we denote with boldface capital letters. If $\vec{\mathbb{P}} = \langle \mathbb{P}_0, \ldots, \mathbb{P}_n \rangle$ is a finite sequence of players symbols, we write $\vec{\mathbb{P}}\mathbb{P}$ for the extension of the sequence by another player symbol $\mathbb{P}$, i.e.,

$$\vec{\mathbb{P}}\mathbb{P} := \langle \mathbb{P}_0, \ldots, \mathbb{P}_n, \mathbb{P} \rangle.$$

A tree T is a finite set of nodes together with a (directed) edge relation (in which any two nodes are connected by exactly one path). Let root_T denote the root of the tree and $\mathrm{tn}(T)$ denote the set of terminal nodes of T. We write $t \in T$ if t is a node in the tree T. If $t \in T$, let $\mathrm{succ}_T(t)$ denote the set of immediate T–successors of t. The DEPTH of the tree T is the number of elements of a longest path in T, and we denote it by $\mathrm{dp}(T)$.

An EXTENSIVE GAME FORM is a tuple $\langle I, T, \mu \rangle$ where I is a set of players, T is a tree and μ is a MOVING FUNCTION. That is,

$$\mu : T \backslash \mathrm{tn}(T) \to I,$$

where $\mu(t) = \mathbb{P}$ indicates that it is $\mathbb{P}$'s move at node t. We call total orders $\succeq$ on $\mathrm{tn}(T)$ PREFERENCES. We fix in advance a finite set $\mathcal{P}$ of preferences to be considered in the formal analysis. A map $\succeq\, : I \to \mathcal{P}$ is called a PREFERENCE PROFILE or DESCRIPTION and we write $\succeq_\mathbb{P}$ for $\succeq(\mathbb{P})$. If $\langle I, T, \mu \rangle$ is an extensive game form, and $\succeq$ is a preference profile (if $t_1 \succeq_i t_2$, we say "player i prefers the node t_1 over node t_2"), then we call

$$\langle I, T, \mu, \succeq \rangle$$

an EXTENSIVE GAME. This model of a game is completely standard and discussions can be found in any game theory text (for example, *cf.* [OR94]).

From now on, we fix an extensive game form $\mathfrak{G} = \langle I, T, \mu \rangle$ and a finite set of preferences $\mathcal{P}$. We consider functions

$$S : T \times I^{\leqslant \mathrm{dp}(T)} \to \mathcal{P}^I,$$

i.e., functions that assign a description $S(t, \vec{\mathbb{P}})$ to a node t in the tree and a finite string $\vec{\mathbb{P}}$ of player symbols (such as $\mathbb{A}\mathbb{B}\mathbb{A}\mathbb{C}$ or also $\varnothing$), and call them STATES.[3]

We now interpret the description $S(t, \varnothing)$ as the TRUE STATE OF AFFAIRS at position t. Note that the true state of affairs can change—for instance, if the game reaches a certain position, players may change their preferences based on what happened in the game so far.

If $S(t, \vec{\mathbb{P}})$ is one of the descriptions defined by the state S, we interpret $S(t, \mathbb{P}\vec{\mathbb{P}})$ as player $\mathbb{P}$'s belief about $S(t, \vec{\mathbb{P}})$. For instance, $S(t, \mathbb{A})$ describes player

---

[3]The restriction of the length of the sequences to a number $\leqslant \mathrm{dp}(T)$ is not essential here. It is just to ensure that states are finite objects. Our algorithm will never need the values of states for longer strings, so we can safely cut the strings at this length.

$\mathbb{A}$'s belief about the true state of affairs at t, and $S(t, \mathbb{A}\mathbb{B})$ describes player $\mathbb{A}$'s belief about player $\mathbb{B}$'s beliefs about the true state.

Note that this way, a state can be easily relativized to the subjective beliefs of a particular player. If S is a state and $\mathbb{P}$ is one of the players, we can define

$$S_{\mathbb{P}}(t, \vec{\mathbb{P}}) := S(t, \mathbb{P}\vec{\mathbb{P}}).$$

The result $S_{\mathbb{P}}$ is essentially a state.[4] A convenient way of presenting a state S (since it presents the belief change in a particularly perspicuous way) is by drawing the corresponding DESCRIPTION TREES $t \mapsto S(t, \vec{\mathbb{P}})$ (for non–terminal nodes t); this is the way we shall present the state in our example in §4.

Our notion of state is very abstract in that it does not take into account any commonsense properties of belief or rationality. For instance, our definition of state allows that $S(t, \varnothing)(\mathbb{A}) \neq S(t, \mathbb{A})(\mathbb{A})$, *i.e.*, player $\mathbb{A}$ has preferences that he doesn't believe he has; or $S(t, \mathbb{A})(\mathbb{P}) \neq S(t, \mathbb{A}\mathbb{A})(\mathbb{P})$, *i.e.*, a violation of positive introspection; or $S(t, \mathbb{A})(\mathbb{B}) = \sqsupseteq$, for some position t inconsistent with $\mathbb{B}$ having the preference $\sqsupseteq$. All of these counterintuitive properties might be conceivably useful in some applications, but can easily be excluded by additional axioms if they are not. For instance, if S is a state, we could call a player $\mathbb{P}$ INTROSPECTIVE in S if for all sequences $\vec{\mathbb{P}}$, we have $S_{\mathbb{P}\vec{\mathbb{P}}} = S_{\mathbb{P}\mathbb{P}\vec{\mathbb{P}}}$, and we can call a state S INTROSPECTIVE if all players are introspective in S. Another useful property is that of egocentricity. In many applications, the agents assume that their beliefs are common belief of all players. This can be formalized as follows: a player $\mathbb{P}$ is called EGOCENTRIC in S if for all sequences $\vec{\mathbb{P}}$, we have $S_{\mathbb{P}} = S_{\mathbb{P}\vec{\mathbb{P}}}$, and we call a state S EGOCENTRIC if all players are egocentric in S.

If $S(t, \vec{\mathbb{P}}) \neq S_{\mathbb{P}}(t, \vec{\mathbb{P}})$, this is an instance of MISTAKEN BELIEF; if t' is a successor of t in the tree T and we have $S_{\mathbb{P}}(t, \vec{\mathbb{P}}) \neq S_{\mathbb{P}}(t', \vec{\mathbb{P}})$, we call it an instance of CHANGING BELIEF. Even though we shall not use it in our example in §4, we can also handle changing preferences: if $S(t, \varnothing)(\mathbb{P}) \neq S(t', \varnothing)(\mathbb{P})$, then player $\mathbb{P}$ CHANGED PREFERENCES between t and t'. If for all $t, t' \in T$ and $\mathbb{P}$, we have $S(t, \vec{\mathbb{P}}) = S_{\mathbb{P}}(t', \vec{\mathbb{P}})$, we call S a STATE OF CORRECT AND UNCHANGING BELIEFS which is an important special case, as it reconstructs the usual Zermelo backward induction solution (cf. Proposition 1).

If $\mathfrak{G} = \langle I, T, \mu \rangle$ is an extensive game form, a GAME MODEL is a tuple $\langle \mathfrak{G}, S \rangle$ where S is a state. Given a game model $\langle \mathfrak{G}, S \rangle$, we can now fully analyze the game and predict its outcome (assuming that the players follow the backward induction solution). In order to do this analysis, we shall construct labellings $\ell_{S_{\vec{\mathbb{P}}}} : T \to tn(T)$ where $\ell_{S_{\vec{\mathbb{P}}}}$ is interpreted as the subjective belief relative to $\vec{\mathbb{P}}$ of the outcome of the game if it has reached the node t. For instance, if $\ell_{S_{\mathbb{A}}}(t) = t^* \in tn(T)$, then player $\mathbb{A}$ believes that if the game reaches t, the eventual outcome is t^*.

---

[4]The only difference is that it is only defined for sequences of length $\leqslant dp(T) - 1$. This is due to our technical restriction discussed in Footnote 3.

THE LABELLING ALGORITHM.   If t is a terminal node, we just let $\ell_U := t$ for all states U. In order to calculate the label of a node t controlled by player $\mathbb{P}$, we need the $\mathbb{P}$–subjective labels of all of its successors. More precisely: If $t \in T$, $\mu(t) = \mathbb{P}$ and we fix a state U, then we can define ℓ_U as follows: find the U–true preference of player $\mathbb{P}$, i.e., $\sqsupseteq = U(t, \varnothing)(\mathbb{P})$. Then consider the labels $\ell_{U_\mathbb{P}}(t')$ for all $t' \in \mathrm{succ}(t)$ and pick the $\sqsupseteq$–maximal of these, say, t^*. Then $\ell_U(t) := t^*$. Concisely, $\ell_U(t)$ is the $U(t, \varnothing)(\mu(t))$–maximal element of the set $\{\ell_{U_{\mu(t)}}(t')\,;\, t' \in \mathrm{succ}(t)\}$.

COMPUTING THE TRUE RUN OF THE GAME.   After we have defined all subjective labellings, the true run can be read off recursively. Since our labels are the terminal nodes, for each t with $\mu(t) = \mathbb{P}$ and S, there is a unique $t' \in \mathrm{succ}(t)$ such that $\ell_{S_\mathbb{P}}(t') = \ell_S(t)$. Starting from the root, take at each step the unique successor determined by $\ell_S(t)$ until you reach a terminal node. Note that this procedure will not follow the subjective tree labellings: player $\mathbb{P}$ might believe that the outcome of the game is t^* when the game reaches t', but the next move is determined by player $\mathbb{P}^* = \mu(t')$ who will use his or her own subjective labelling $\ell_{S_{\mathbb{P}^*}}$ in order to determine the next successor.

PROPOSITION 1 *If S is a state of correct and unchanging beliefs, then our algorithm reproduces the usual backward induction solution.*

Proof: An easy inductive argument shows that if S is a state of correct and unchanging beliefs, then for all $\vec{\mathbb{P}}$, we have $\ell_S = \ell_{S_{\vec{\mathbb{P}}}}$. We also know that for any $\mathbb{P}$, t and $U = S_{\vec{\mathbb{P}}}$, $U(t, \varnothing)(\mathbb{P})$ is the true preference of player $\mathbb{P}$, say, $\succeq_\mathbb{P}$.

Thus, for any non–terminal node t, $\ell_U(t)$ is the $\succeq_{\mu(t)}$–maximal element of the set $\{\ell_U(t')\,;\, t' \in \mathrm{succ}(t)\}$. This is exactly the definition of the backward induction labelling. $\qquad\square$

PROPOSITION 2 *If S is a state and $\mathbb{P}$ is egocentric in S, then for all sequences $\vec{\mathbb{P}}$, we have that $\ell_{S_\mathbb{P}} = \ell_{S_{\mathbb{P}\vec{\mathbb{P}}}}$.*

Proof: We prove $\ell_{S_\mathbb{P}}(t) = \ell_{S_{\mathbb{P}\vec{\mathbb{P}}}}(t)$ by induction on the depth of the node t. For terminal nodes t, the claim is true by definition. If t is nonterminal, then $\ell_{S_\mathbb{P}}(t)$ is the $S_\mathbb{P}(t, \varnothing)(\mu(t))$–maximal element of the set $\{\ell_{S_{\mathbb{P}_{\mu(t)}}}(t')\,;\, t'$ is an immediate successor of $t\}$, and $\ell_{S_{\mathbb{P}\vec{\mathbb{P}}}}(t)$ is the $S_{\mathbb{P}\vec{\mathbb{P}}}(t, \varnothing)(\mu(t))$–maximal element of the set $\{\ell_{S_{\mathbb{P}\vec{\mathbb{P}}_{\mu(t)}}}(t')\,;\, t'$ is an immediate successor of $t\}$. But since $\mathbb{P}$ is egocentric, we have $S_\mathbb{P}(t, \varnothing)(\mu(t)) = S_{\mathbb{P}\vec{\mathbb{P}}}(t, \varnothing)(\mu(t))$ and $\ell_{S_{\mathbb{P}_{\mu(t)}}}(t') = \ell_{S_{\mathbb{P}\vec{\mathbb{P}}_{\mu(t)}}}(t') = \ell_{S_\mathbb{P}}(t')$ by induction hypothesis, and thus the sets of labels of successors are identical. $\quad\square$

4   ANALYSIS OF THE STORY

In the following, we now transform the story presented in §2 into a state that our formal system can deal with. Actually, we do not need to give a full state but can restrict ourselves to those values of the function S that are relevant for the algorithm given in §3. Note that this analysis cannot be done in an
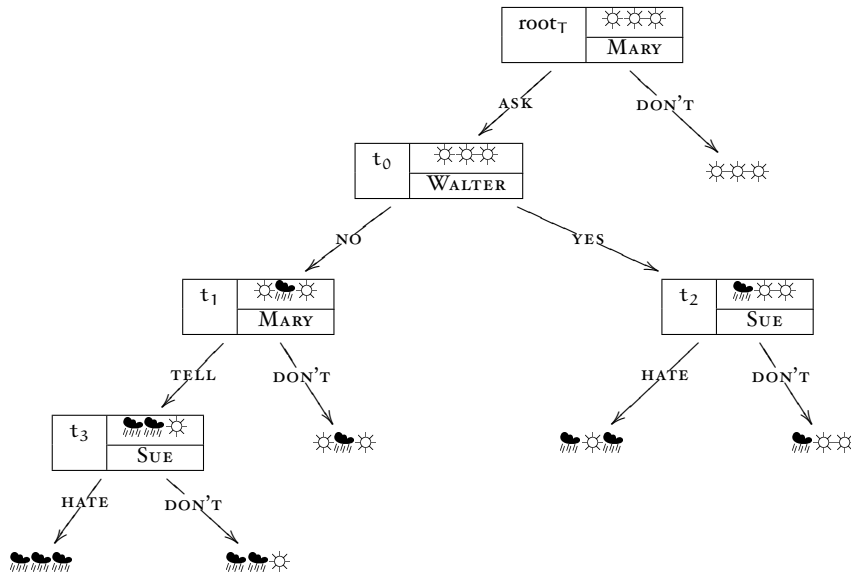
Figure 1: The game tree for the story about Walter, Mary and Sue.

automated way: a human theorist is needed to transform the natural language story into its formal counterpart. The natural language of the story has to be evaluated into categories of 'relevant' and 'irrelevant' for the plot line. For instance, there is a potential agent in our story, Jeff, who is structurally irrelevant for the outcome. Also, there are a number of statements that could indicate choices of the agents ("*One morning, she gives Walter an ultimatum…*": could she have done it in the evening? "*…he used his cell phone to call Mary…*": could he have used a pay phone?), and the human modeller will have to decide whether they are relevant for the plot or not. In general, this is a subtle decision process that could require a thorough analysis of the text of the story.

In our story, there are three relevant agents $\{\mathbb{W}, \mathbb{M}, \mathbb{S}\}$. By analyzing the decision points in the story line (*i.e.*, those moments in which one of the agents makes a decision about an action), we can see that there are five of those, corresponding to the non–terminal nodes in our game tree. We see easily that there are six conceivably possible outcomes in the story (the terminal nodes in the game tree). The game tree that results from this analysis of the story is given in Figure 1. In order to have an easily recognizable notation for the six possible outcomes, we use the notation ☼ to indicate that a relationship is intact and ☔ to indicate that it is ended, and list the terminal nodes by the status of the relationships in the order Sue–Walter, Mary–Walter and Mary–Sue; for instance, ☔☼☔ stands for "Walter left Sue, is together with Mary, and Sue hates Mary". Since each outcome uniquely identifies one of the terminal nodes, we shall refer to the terminal nodes by these labels.
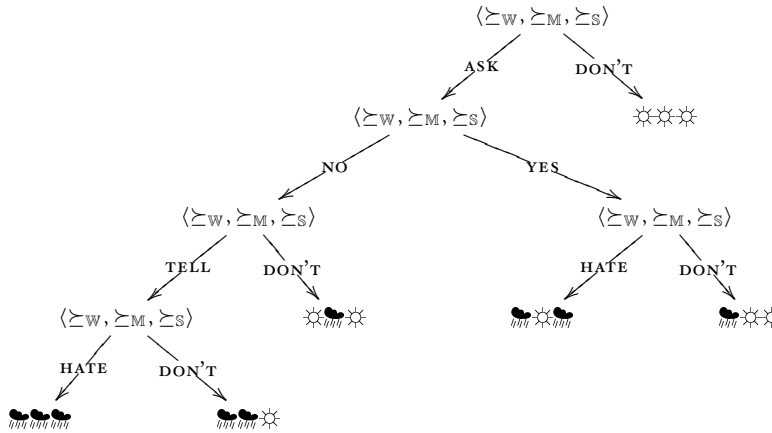
Figure 2: The description tree $t \mapsto S(t, \varnothing)$ of true preferences

By analyzing the decision procedures in the story further, we realize that there are only five preferences relevant for the analysis of the story: the true preferences of all three players, the mistaken beliefs of Walter about Sue's preferences at the beginning of the game, and the mistaken beliefs of Mary about Walter's preferences at the beginning of the game. We denote these preferences by $\succeq_W$, $\succeq_M$, $\succeq_S$, $\succeq_{S*}$, and $\succeq_{W*}$, respectively:
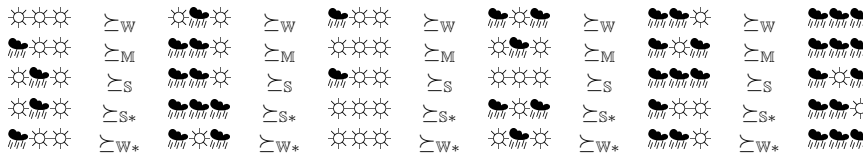
In our situation, a description is a triple of preferences representing the preferences of Walter, Mary, and Sue. For instance, the true description at root_T is the triple

$$S(\mathrm{root}_T, \varnothing) := \langle \succeq_W, \succeq_M, \succeq_S \rangle.$$

As mentioned before, we shall give the state S mostly in terms of the description trees. We start with the true state of affairs in Figure 2. Note that in this tree, the description at each node is the same. This corresponds to the fact that there are no changing preferences in our example.

We continue with Walter's beliefs. Walter's first order belief is just $\langle \succeq_W, \succeq_M, \succeq_{S*} \rangle$ at each node: he is wrong about Sue, but correct about Mary. The belief revision that Walter undergoes in the story happens on the level of his second order beliefs: we give the description tree for S_{WM} in Figure 3 showing his belief revision at stage t_0. He was originally egocentric, as $S_{WM}(\mathrm{root}_T, \varnothing) = S_W(\mathrm{root}_T, \varnothing)$, but realizes after Mary plays ASK that Mary cannot be right in her predictions of his own actions. He therefore corrects
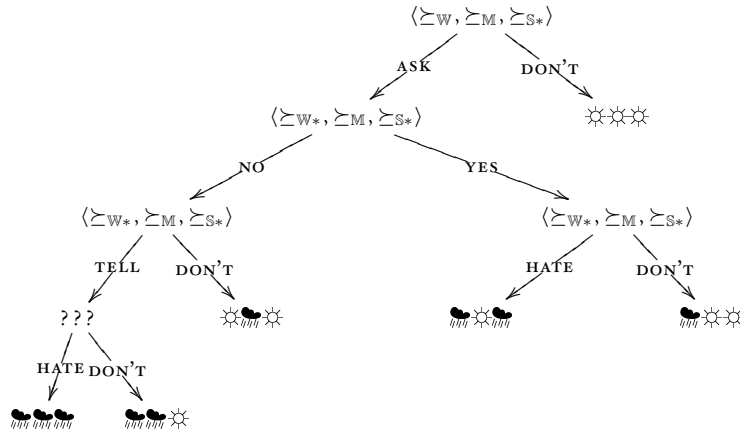
Figure 3: The description tree $t \mapsto S(t, \mathbb{WM}) = S_{\mathbb{WM}}(t, \varnothing)$ of Walter's beliefs about Mary's beliefs. The beliefs that cannot be extracted from the story are marked by "? ? ?".

his beliefs about her beliefs to $\succeq_{W*}$ (which is what Mary actually believes). Unfortunately for him, he is still wrong about Mary's belief about Sue.[5] Note that we do not give a value for $S(t_3, \mathbb{WM})$: abstractly, we know that Walter must undergo yet another belief revision here, as Mary's action TELL is inconsistent with the belief that Mary acts according to the preference $\succeq_M$ and the belief that Sue according to $\succeq_{S*}$. However, in the story, we just learn that Walter is confused ("*Walter gazed emptily at his paintings searching for a logician to help him to figure out what had happened*") and are thus not able to reconstruct a concrete belief revision for Walter at this step. Of course, since Walter does not move after t_3, his beliefs are not relevant for the outcome of the game anymore.

The final description tree is that of Mary in Figure 4. She starts off with correct beliefs about Sue and false beliefs about Walter. After Walter plays NO in t_0, Mary corrects her belief about Walter. For the calculation of the subjective labellings, we furthermore need the states S_{MW} and S_{MWM}. The story suggests that Mary has no complicated iterated beliefs and that neither Walter nor Mary fail introspection. Therefore, we assume that Mary is egocentric and that Walter is introspective in the sense of § 3. Therefore, $S_{MW} = S_{MWM} = S_M$.

Based on the description trees of S, S_M, S_W, S_{MW}, S_{WM}, and S_{MWM}, we shall now proceed to compute the values of $\ell_U(t)$ for the relevant states U. Under an additional assumption of weakened egocentricity (see STEP 5), we can calculate some additional values.

---

[5]The fact that Walter can be right about Mary's preferences but wrong about his beliefs about what she'll do is a crucial point in our algorithm. Knowing Mary's preferences does not allow Walter to predict what she will do: since he is wrong about Sue's preferences and (more importantly) about Mary's beliefs about Sue's preferences, he cannot correctly predict the outcome of the game in Mary's subjective labelling.
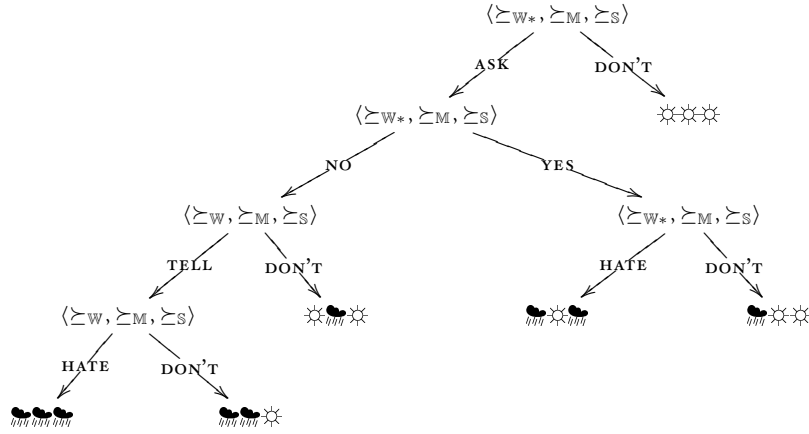
Figure 4: The description tree $t \mapsto S(t, \mathbb{M}) = S_\mathbb{M}(t, \varnothing)$ of Mary's beliefs. This is also the description tree of $S_{\mathbb{MW}}$ and $S_{\mathbb{MWM}}$.

Step 1. *Terminal nodes.* By definition, for all states U, $\ell_U(t) = t$ if $t \in tn(T)$.

Step 2. *Nodes of depth one.* There are two nodes of depth one, viz. t_2 and t_3. In these cases, we can simply exploit the recursive definition and Step 1 to get the following characterizations:

$$\ell_U(t_2) = \begin{cases} \text{🐜☀☀} & \text{if } U(t_2, \varnothing)(\mathbb{S}) = \succeq_\mathbb{S} \\ \text{🐜☀🐜} & \text{if } U(t_2, \varnothing)(\mathbb{S}) = \succeq_{\mathbb{S}*}, \end{cases}$$

$$\ell_U(t_3) = \begin{cases} \text{🐜🐜☀} & \text{if } U(t_3, \varnothing)(\mathbb{S}) = \succeq_\mathbb{S} \\ \text{🐜🐜🐜} & \text{if } U(t_3, \varnothing)(\mathbb{S}) = \succeq_{\mathbb{S}*} . \end{cases}$$

In particular, we get the following values for our six states under consideration:

$$\left.\begin{matrix} \ell_S(t_2) \\ \ell_{S_\mathbb{M}}(t_2) \\ \ell_{S_\mathbb{MW}}(t_2) \\ \ell_{S_\mathbb{MWM}}(t_2) \end{matrix}\right\} = \text{🐜☀☀}, \qquad \left.\begin{matrix} \ell_S(t_3) \\ \ell_{S_\mathbb{M}}(t_3) \\ \ell_{S_\mathbb{MW}}(t_3) \\ \ell_{S_\mathbb{MWM}}(t_3) \end{matrix}\right\} = \text{🐜🐜☀},$$

$$\left.\begin{matrix} \ell_{S_\mathbb{W}}(t_2) \\ \ell_{S_\mathbb{WM}}(t_2) \end{matrix}\right\} = \text{🐜☀🐜}, \qquad \left.\begin{matrix} \ell_{S_\mathbb{W}}(t_3) \\ \ell_{S_\mathbb{WM}}(t_3) \end{matrix}\right\} = \text{🐜🐜🐜}.$$

Step 3. *Nodes of depth two.* The node t_1 is the only node of depth two and it is a node controlled by Mary. Therefore, in order to calculate $\ell_U(t_1)$, we need the labels of the immediate successors of t_1 in the labelling $\ell_{U_\mathbb{M}}$. The immediate successors of t_1 are ☀🐜☀ and t_3. By Step 1, $\ell_U(\text{☀🐜☀}) = \text{☀🐜☀}$ for all U. By Step 2, we know the value of $\ell_{U_\mathbb{M}}(t_3)$ for $U \in \{S, S_\mathbb{W}, S_\mathbb{MW}\}$. A glance at the table in Step 2 gives us the values

$$\ell_{S_M}(t_3) = \ell_{S_{MWM}}(t_3) \;=\; \text{👣👣☼},$$
$$\text{and } \ell_{S_{WM}}(t_3) \;=\; \text{👣👣👣}.$$

So, in the first two cases, we have to compare ☼👣☼ and 👣👣☼ according to the preference that the state U assigns to Mary. We see that $S(t_1, \varnothing)(\mathbb{M}) = S_{MW}(t_1, \varnothing) = \succeq_M$ and get

$$\ell_S(t_1) = \ell_{S_{MW}}(t_1) = \text{👣👣☼}.$$

In the case of S_W, we have to compare 👣👣👣 and ☼👣☼ according to $\succeq_M = S_W(t_1, \varnothing)(\mathbb{M})$, and get
$$\ell_{S_W}(t_1) = \text{☼👣☼}.$$

Assuming Mary's introspection, we can use $\ell_{MM}(t_3) = \text{👣👣☼}$ and calculate $\ell_M(t_1) = \text{👣👣☼}$.

STEP 4. *Nodes of depth three.* The only node of depth three is the node t_0 controlled by Walter. Its immediate successors are t_1 and t_2, and thus we need $\ell_{U_W}(t_1)$ and $\ell_{U_W}(t_2)$ for a state U in order to calculate $\ell_U(t_0)$. Checking the calculations of STEP 2 (for t_2) and STEP 3 (for t_2), we see that we have determined these labels for $U \in \{S, S_M\}$, and received the values

$$\ell_{S_W}(t_1) \;=\; \text{☼👣☼}, \quad \ell_{S_{MW}}(t_1) \;=\; \text{👣👣☼},$$
$$\ell_{S_W}(t_2) \;=\; \text{👣☼👣}, \quad \ell_{S_{MW}}(t_2) \;=\; \text{👣☼☼}.$$

For $U = S$, we compare ☼👣☼ and 👣☼👣 according to $\succeq_W$ and get $\ell_S(t_0) = \text{☼👣☼}$, for $U = S_M$, we compare 👣👣☼ and 👣☼☼ according to $\succeq_{W*}$ and get $\ell_{S_M}(t_0) = \text{👣☼☼}$.

Using Walter's introspection (and thus using $\ell_{S_{WW}}(t_1) = \text{☼👣☼}$ and $\ell_{S_{WW}}(t_2) = \text{👣☼👣}$), we can also calculate $\ell_{S_W}(t_0) = \text{☼👣☼}$.

STEP 5. *The root.* The root is a node controlled by Mary, so we need to compare $\ell_{U_M}(t_0)$ and $\ell_{U_M}(\text{☼☼☼}) = \text{☼☼☼}$. We only have $\ell_{S_M}(t_0) = \text{👣☼☼}$, so we check that $\text{👣☼☼} \succeq_M \text{☼☼☼}$, and get $\ell_S(\mathrm{root}_T) = \text{👣☼☼}$.

As before, under the assumption of Mary's introspection, we get that $\ell_{S_{MM}}(t_0) = \text{👣☼☼}$, and thus $\ell_{S_M}(\mathrm{root}_T) = \text{👣☼☼}$.

To calculate $\ell_{S_W}(\mathrm{root}_T)$, we need to make the additional (but very mild and realistic) weakened egocentricity assumption of $S_{WM\bar{P}} = S_{WM}$. Then, going through STEP 1 to STEP 4 again, we can calculate $\ell_{S_{WMWM}}(t_3) = \text{👣👣👣}$, $\ell_{S_{WMW}}(t_2) = \text{👣☼👣}$, $\ell_{S_{WMW}}(t_1) = \text{👣👣👣}$, $\ell_{S_{WM}}(t_0) = \text{👣☼👣}$, and finally $\ell_{S_W}(\mathrm{root}_T) = \text{☼☼☼}$.

We finished our calculation of labels. It it useful to note again that the labelling functions behave quite differently from the usual backward induction labellings. In backward induction with common knowledge of the true preferences (in the notation of §3, this means a state S of correct and unchanging
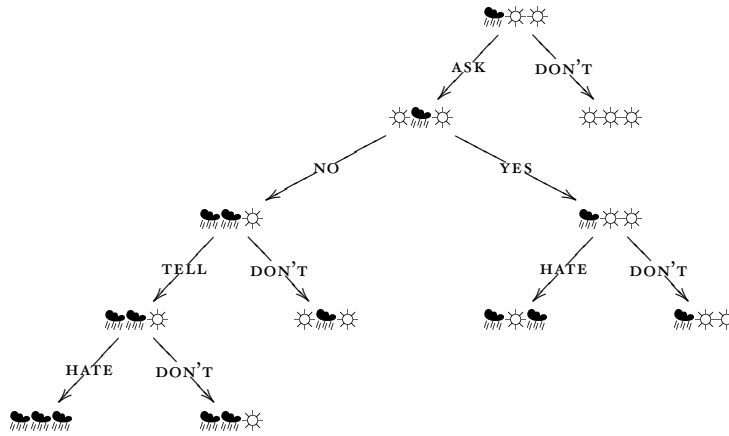
Figure 5: The labelling $t \mapsto \ell_S(t)$.

beliefs), you get one labelling corresponding to ℓ_S that you can use to read off the moves of the game: we discussed this as a special case of our set-up in Proposition 1. In particular, in this unique labelling each node has at least one successor with the same label. In our case, the ℓ_U–label at t only indicates the belief (relative to a given state U) of the player $\mu(t)$ controlling the node about the eventual outcome. As a consequence, we lose the direct connection between $\ell_U(t)$ and $\ell_U(t')$ for $t' \in \text{succ}(t)$. Since $\ell_U(t)$ corresponds to $\mu(t)$'s beliefs, the value of $\ell_U(t)$ is connected to $\ell_{U\mu(t)}(t')$ instead.

Consider the labelling ℓ_S in Figure 5 where the nodes root_T and t_0 have no successor nodes with the same label. For instance, in order to understand $\ell_S(t_0)$, we have to look at the labelling ℓ_{S_W} (as given in Figure 6), realizing that the value $\ell_S(t_0) = $ ☼🐾☼ corresponds to the $\succeq_W$–maximal label among the ℓ_{S_W}–values of the successors of t_0.

We shall now use the labellings calculated in order to give an interpretation of the story and explain the course of events. We start in the node root_T which is a node in which Mary has to play, and consider Mary's subjective labelling ℓ_{S_M} as given in Figure 7. Mary believes that Walter will answer YES to her ultimatum, corresponding to "*She is convinced that if she presses Walter enough, he will finally leave Sue for her.*" In Figure 7, this is represented by $\ell_{S_M}(t_0) = $ 🐾☼☼. Mary's belief that "*she can make up with Sue afterwards*" is represented by $\ell_{S_M}(t_3) = $ 🐾🐾☼ and $\ell_{S_M}(t_2) = $ 🐾☼☼. The label $\ell_S(\text{root}_T) = $ 🐾☼☼ tells us that Mary plays ASK in her first move ("*One morning, she gives Walter an ultimatum: he should make up his mind and choose between her and Sue.*").

In the story, we can now read the deliberations of Walter. He realizes that he had wrong beliefs about Mary ("*Walter is ultimately confused: he must have misjudged Mary*") and revises his belief as discussed above. He now reasons about Mary's actions if he plays NO as the next move: "*if he said no to Mary, would she*
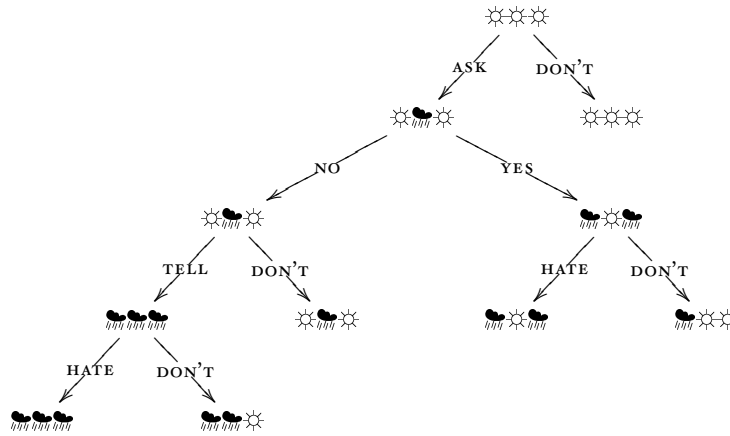
Figure 6: The (partial) labelling $t \mapsto \ell_{S_W}(t)$. The label at t_0 has been computed under the assumption of introspection; the label at root_T has been computed with the additional assumption of $S_{WM\bar{P}} = S_{WM}$ (cf. STEP 5).

tell Sue? No, he reasoned, since then she would lose both Sue and him which is definitely worse than just losing him." corresponds to $\ell_{S_W}(t_1) = $ ☼🐾☼. As a consequence, Walter plays NO in the second move ("*Smiling, he used his cell phone to call Mary and tell her that he would not leave Sue*").

The third move is now Mary's again who revises her beliefs about Walter ("*Well, if that's what he wants, then I must have completely misjudged him.*") She follows her preferences and beliefs about Sue's preferences: "*I should cut my losses, and at least be honest to my best friend and tell her,*" she reasoned, and acted accordingly." Mary plays TELL in the third move.

Finally, Sue follows Mary's predictions and plays DON'T: "*And Mary was right in her judgement of Sue. The two women discussed the matter, and when Walter returned from his appointment in the afternoon, his paintings were standing on the front lawn of their house and the lock of the front door had been changed.*".

As part of our formal analysis, we can now also go beyond describing the behaviour of agents. For instance, the fact that Walter is surprised by Mary's action ASK is represented by $\ell_{S_W}(\text{root}_T) = $ ☼☼☼ ("*Walter convinced himself that Mary will never ask him to separate from Sue*"). We can also do counterfactual reasoning about the storyline in terms of our labelling. This would become relevant for programming the reasoning of artificial agents based on our system (cf. § 5.5).

5 CONCLUSION, RELATED WORK AND FUTURE WORK

With the formal analysis of §§ 3 and 4, we fulfilled the goal mentioned at the end of § 2: we have a formal system that allows to mimic the intuitive reasoning of human beings about the game situation. However, the definition of a formal
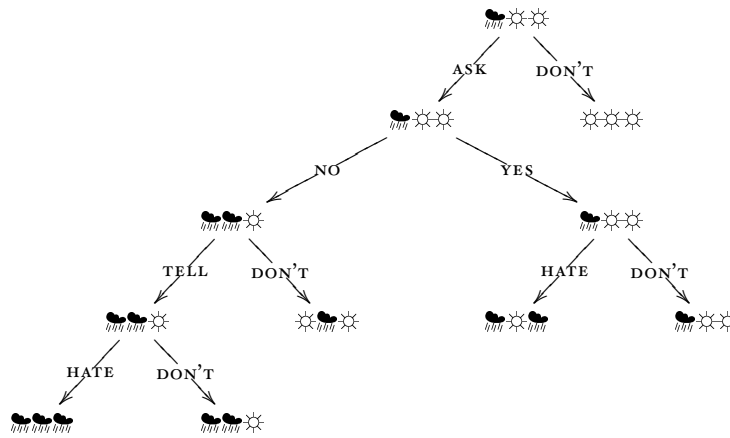
Figure 7: The (partial) labelling $t \mapsto \ell_{S_M}(t)$. The value at t_1 has been computed under the assumption of introspection as in Step 3.

system provides only the very first step. A lot of open questions and problems remain.

5.1 RELATED WORK

There are many other models for game–theoretic reasoning in the presence of uncertainty, and it would be wrong to claim that they cannot deal with the example given in this paper at least as well as our abstract framework. It should be stressed that the elaborate example in this paper is not more than that: an example to show how our algorithm works; we do not claim that our approach is the first one that can explain the behaviour of the agents in stories like this. We also do not claim that our algorithm adds to our understanding of game theory in general: game–theoretically, it is just the straightforward application of the basic idea of backward induction to the situation of changing and mistaken beliefs.

The benefit of our framework is its extraordinary simplicity: we make the player's preferences the basic entities of the entire algorithm and encode the belief change into the notion of *state*, thus avoiding to have to discuss the belief change functions. Because of this, we get a very parsimonious and flexible algorithm that can be applied to many different situations.

Closest in spirit to our game–theoretic analysis is the excellent paper [Fei05] in which he provides a powerful language to talk about dynamic games. Clearly, Feinberg's system is able to provide a clear analysis of our example story. Feinberg does provide notions of rationality for his system and is thus able to *derive* that the agents will follow the backward induction strategy under certain conditions [Fei05, § 4.2]. This relates closely to our § 5.2.

Our parsimonious language suggests an affinity with the dynamic logic solutions of logical puzzles. A number of modal logic approaches to reasoning about belief revision and preferences exist; among these, [vDL07], the *epistemic preference models* of [vBL07] (based on Hansson's [Han90] logics of preferences) and the proba–doxastic logic of [Auc05a, Auc05b] come closest to our approach. We discuss this relationship more closely in § 5.3.

5.2   GAME–THEORETIC PROBLEMS.

Our analysis presupposes backward induction. In order to connect our work to the large body of research on the foundations of rationality mentioned in § 1, one could turn this around by making descriptions tuples of *strategies* instead of preferences (i.e., $S(t, \mathbb{P})(\mathbb{Q}) = \Sigma$ meaning that in state S, player $\mathbb{P}$ believes that player $\mathbb{Q}$ follows the strategy Σ), and then analyze properties of the players (corresponding to rationality) that will ensure that players follow the subjective backward induction strategy. If we build a logical language on top of our system with some expressive power, we should make our system resemble Feinberg's *subjective framework* from [Fei05]. Under appropriate assumption, it is likely that analogues of Feinberg's *subjective conditions on rationality implying backward induction* can be found for our system and analogues of Aumann's classic theorem [Aum95] and Feinberg's [Fei05, Proposition 10] can be proved.

5.3   THE RELATIONSHIP TO DYNAMIC EPISTEMIC LOGIC.

As indicated in § 5.1, the purely combinatorial character of our algorithm suggests a very close relation to the techniques of dynamic epistemic logic, and it is likely that there is a translation of our algorithm into the framework of dynamic epistemic logic.

Since backward induction can be expressed in dynamic epistemic logic [vB01] and our algorithm is just based on a dynamic doxastic system and backward induction, it should be possible to do for our algorithm what van Ditmarsch did to belief revision in his [vD05].

5.4   ANALYSIS OF STORY LINES.

A conceptual system using our formal analysis could be used to analyze story lines. The question of what constitutes an interesting story has become more interesting for computer scientists as part of game design. Obviously, a large part of the entertainment that people get out of watching crime stories on TV or playing story–based games is unrelated to the actual plot of the stories, and rests in properties superficial to the actual story (such as the graphics animation in computer games or the quality of the actor in a TV drama). However, we believe that at least part of the entertainment derives from interesting epistemic story twists, either for the viewer or for the characters in the story. This

type of analysis has been implemented in the *Story Understanding* community [Rum75, Leh81, Mue08].

Our simple formal analysis provides an abstract means of classifying story-lines based on their epistemic and doxastic properties that could be more perspicuous and easy to handle than the known *Story Understanding* systems. An empirical project based on this would be to analyze a large sample of stories (e.g., plots of crime stories) using our framework and find out whether certain formal patterns occur more frequently than others. For a first step towards this goal, cf. [LPS08] in which the authors investigate ten narratives from the series *CSI: Crime Scene Investigation*™ and identify a small number of formal building blocks using the system of this paper.

5.5 GAME DESIGN.

The algorithm given in this paper could be used by artificial agents in computer games to mimic human behaviour. Epistemically more complicated computer games could feature artificial agents with more involved goals (such as trying to bring the human player to perform a certain action or trying to get a certain piece of information from the human player without the player being aware of this) that requires them to assess the doxastic situation.

An alternative use of our algorithm in game design could be connected to the project described above: if our formal model allows a classification of those epistemic story lines that human audiences find interesting as discussed in §5.4, this information could be used to reverse engineer story lines (for instance for auto–generation of subplots in a computer game) in order to avoid story lines that are too straightforward or too complicated for the human players to enjoy.

REFERENCES

[Auc05a]   GUILLAUME AUCHER. A combined system for update logic and belief revision. In MICHAEL W. BARLEY and NIK KASABOV, editors, *Intelligent Agents and Multi–Agent Systems, 7th Pacific Rim Workshop on Multi–Agents,* PRIMA 2004, *Auckland, New Zealand, August 8–13, 2004, Revised Selected Papers*, volume 3371 of *Lecture Notes in Artificial Intelligence*, pages 1–17. Springer–Verlag, 2005.

[Auc05b]   GUILLAUME AUCHER. How our beliefs contribute to interpret actions. In MICHAL P. PETTA and LASZLO Z. VARGA, editors, *Multi–Agent Systems and Applications* IV, *4th International Central and Eastern European Conference on Multi-Agent Systems,* CEEMAS *2005, Budapest, Hungary, September 15–17, 2005, Proceedings*, volume 3690 of *Lecture Notes in Artificial Intelligence*, pages 276–286. Springer–Verlag, 2005.

[Aum95]   ROBERT AUMANN. Backward induction and common knowledge
            of rationality. *Games and Economic Behavior*, 8:6–19, 1995.

[Aum99]   ROBERT AUMANN. Interactive epistemology I: Knowledge. *International Journal of Game Theory*, 28:263–300, 1999.

[BB99]   GIACOMO BONANNO and PIERPAOLO BATTIGALLI. Recent results on belief, knowledge and the epistemic foundations of game
            theory. *Research in Economics*, 53:149–225, 1999.

[BDDM04]   JAN BROERSEN, FRANK DIGNUM, VIRGINIA DIGNUM, and
            JOHN-JULES MEYER. Designing a deontic logic of deadlines. In
            ALESSIO LOMUSCIO and DONALD NUTE, editors, *Deontic Logic
            in Computer Science, 7th International Workshop on Deontic Logic in
            Computer Science,* DEON *2004, Madeira, Portugal, May 26-28, 2004,
            Proceedings*, volume 3065 of *Lecture Notes in Artificial Intelligence*,
            pages 43–56. Springer–Verlag, 2004.

[Bin96]   KEN BINMORE. A note on backward induction. *Games and Economic Behavior*, 17:135–137, 1996.

[Bin97]   KEN BINMORE. Rationality and backward induction. *Journal of
            Economic Methodology*, 4:23–41, 1997.

[Boa04]   OLIVER BOARD. Dynamic interactive epistemology. *Games and
            Economic Behavior*, 49:49–80, 2004.

[Bra07]   ADAM BRANDENBURGER. The power of paradox: some recent
            developments in interactive epistemology. *International Journal of
            Game Theory*, 35:465–492, 2007.

[BS03]   PIERPAOLO BATTIGALLI and MARCIANO SINISCALCHI. Rationalization and incomplete information. *Advances in Theoretical
            Economics*, 3:Article 3, 2003.

[dB04]   BOUDEWIJN DE BRUIN. *Explaining Games: On the Logic of Game
            Theoretic Explanations*. PhD thesis, Universiteit van Amsterdam,
            2004. ILLC Publications DS-2004-03.

[Fei05]   YOSSI FEINBERG. Subjective reasoning — dynamic games. *Games
            and Economic Behavior*, 52:54–93, 2005.

[Gär92]   PETER GÄRDENFORS. Belief revision: an introduction. In
            PETER GÄRDENFORS, editor, *Belief Revision*, volume 29 of *Cambridge Tracts in Theoretical Computer Science*, pages 1–28. Cambridge
            University Press, 1992.

[GS53]   DAVID GALE and FRANK M. STEWART. Infinite games with per-
             fect information. IN HAROLD W. KUHN and ALBERT W. TUCKER,
             editors, *Contributions to the Theory of Games* II, volume 28 of *An-
             nals of Mathematical Studies*, pages 245–266. Princeton University
             Press, 1953.

[Han90]   SVEN OVE HANSSON. Preference–based deontic logic (PDL).
             Journal of Philosophical Logic, 19:75–93, 1990.

[Leh81]   WENDY G. LEHNERT. Plot units and narrative summarization.
             Cognitive Science, 4:293–331, 1981.

[LPS08]   BENEDIKT LÖWE, ERIC PACUIT, and SANCHIT SARAF. Analyzing
             stories as games with mistaken and changing beliefs, 2008. ILLC
             Publications PP-2008-31.

[MBGVS97]   PHILIPPE MONGIN, MICHAEL O. L. BACHARACH, LOUIS-
             ANDRÉ GÉRARD-VARET, and HYUN SONG SHIN, EDITORS. *Epis-
             temic logic and the theory of games and decisions*, volume 20 of *Theory
             and Decision Library, Series C: Game Theory, Mathematical Program-
             ming and Mathematical Economics*. Kluwer Academic Publishers,
             1997.

[MP92]   RICHARD D. MCKELVEY and THOMAS R. PALFREY. An experi-
             mental study of the centipede game. *Econometrica*, 60:803–836,
             1992.

[Mue08]   ERIK T. MUELLER. Story understanding. In LYNN NADEL, editor,
             Encyclopedia of Cognitive Science. John Wiley & Sons, Inc., 2008.

[NT98]   ROSEMARIE NAGEL and FANG FANG TANG. An experimental
             study on the centipede game in normal form — an investigation
             on learning. *Journal of Mathematical Psychology*, 42:356–384, 1998.

[OR94]   MARTIN J. OSBORNE and ARIEL RUBINSTEIN. *A course in game
             theory*. MIT Press, Cambridge, MA, 1994.

[Par02]   ROHIT PARIKH. Social software. *Synthese*, 132:187–211, 2002.

[Pea84]   DAVID G. PEARCE. Rationalizable strategic behavior and the
             problem of perfection. *Econometrica*, 52(4):1029–1050, 1984.

[Per]   ANDRÈS PEREA. A model of minimal probabilistic belief revi-
             sion. *Theory and Decision*. to appear.

[Per07]   ANDRÈS PEREA. Epistemic foundations for backward induc-
             tion: An overview. In JOHAN VAN BENTHEM, DOV GABBAY, and
             BENEDIKT LÖWE, editors, *Interactive Logic. Selected Papers from the*

7th Augustus de Morgan Workshop, London, volume 1 of *Texts in Logic and Games*, pages 159–193. Amsterdam University Press, 2007.

[Ros81]   ROBERT W. ROSENTHAL. Games of perfect information, predatory pricing and the chain store. *Journal of Economic Theory*, 25:92–100, 1981.

[Rum75]   DAVID E. RUMELHART. Notes on a schema for stories. In DANIEL G. BOBROW and COLLINS ALLEN M., editors, *Representation and Understanding: Studies in cognitive science*, pages 211–236. Academic Press, 1975.

[Sta98]   ROBERT STALNAKER. Belief revision in games: forward and backward induction. *Mathematical Social Sciences*, 36:31–56, 1998.

[vB01]   JOHAN VAN BENTHEM. Games in dynamic–epistemic logic. *Bulletin of Economic Research*, 53(4):219–248, 2001.

[vBL07]   JOHAN VAN BENTHEM and FENRONG LIU. Dynamic logic of preference upgrade. *Journal of Applied Non–Classical Logic*, 17:157–182, 2007.

[vD00]   HANS P. VAN DITMARSCH. *Knowledge Games*. PhD thesis, Rijksuniversiteit Groningen, 2000. ILLC Publications DS-2000-06.

[vD05]   HANS VAN DITMARSCH. Belief change and dynamic logic. In JAMES DELGRANDE, JÉRÔME LANG, HANS ROTT, and JEAN-MARC TALLON, editors, *Belief Change in Rational Agents: Perspectives from Artificial Intelligence, Philosophy, and Economics, Proceedings of Dagstuhl Seminar 05321, Schloß Dagstuhl*, 2005.

[vDL07]   HANS VAN DITMARSCH and WILLEM LABUSCHAGNE. My beliefs about your beliefs — a case study in theory of mind and epistemic logic. *Synthese*, 155:191–209, 2007.

[Zer13]   ERNST ZERMELO. Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels. In E. W. HOBSON and A. E. H. LOVE, editors, *Proceedings of the Fifth International Congress of Mathematicians, Cambridge 1912, Volume 2*, pages 501–504. Cambridge University Press, 1913.