Christian W. Bach and Conrad Heilmann

Agent connectedness and backward induction

Working paper

Agent Connectedness and Backward Induction

Christian W. Bach[*] and Conrad Heilmann[**]

Abstract. We analyze the sequential structure of dynamic games with perfect information. A three-stage account is proposed, that specifies set-up, reasoning and play stages. Accordingly, we define a player as a set of agents corresponding to these three stages. The notion of agent connectedness is introduced into a type-based epistemic model. Agent connectedness measures the extent to which agents' choices are sequentially stable. Thus describing dynamic games allows to more fully understand strategic interaction over time. In particular, we provide sufficient conditions for backward induction in terms of agent connectedness. Also, our framework makes explicit that the epistemic independence assumption involved in backward induction reasoning is stronger than usually presumed, and makes accessible multiple-self interpretations for dynamic games.

1 Introduction

Dynamic games model sequential strategic interaction. The standard extensive form models dynamic games as trees, but does not further explicate the sequential dimension. The structure of the game, the players, their reasoning and strategies are implicitly assumed to remain stable throughout the whole game. In particular, the reasoning is supposed to occur before the game and to apply to the entire duration of the game. However, local deviations from strategies are relevant for the dynamics of sequential interaction. More specifically, agents may depart from the strategy of their respective player, thus contradicting the idea that agents act according to instructions. Here, we perceive of a player as a set of agents and introduce the notion of *agent connectedness* to capture the extent of sequential stability of players. In our account, high agent connectedness characterizes an agent's compliance with a player and low agent connectedness an agent's deviation. Precisely such properties of agents are central to backward induction, since players need to be able to entertain deviating moves by opponents' agents in hypothetical reasoning. Indeed, here we provide sufficient conditions for backward induction in terms of agent connectedness.

In a general sense, we amend the representation of a dynamic game by three sequential stages. In the set-up stage, the game structure and the players' utilities are determined. Then, in the reasoning stage, the players deliberate about the

[*] University of Lausanne, Faculty of Business and Economics (HEC), Email: christianwoldemar.bach@unil.ch

[**] London School of Economics (LSE), Department of Philosophy, Logic and Scientific Method, Email: c.heilmann@lse.ac.uk

game, their opponents and choose their strategies. Finally, in the play stage, the players' agents act at their respective decision nodes. Relative to these three sequential stages, a player is defined as a set of agents, namely the set-up agent, the reasoning agent and the game agents. We also amend the notion of strategy such that its use in the stages can be discussed separately, introducing the notion of initial strategy in the reasoning stage and actual strategy in the play stage. This three-stage account enables us to make explicit the sequential stability assumptions inherent in dynamic games. Also, the framework can be used to relax such assumptions locally.

The reasoning of players in games is usually described by epistemic models. Here, we extend a type-based epistemic model of dynamic games with an initial strategy function by means of which the connectedness of each agent to his respective player can be expressed. In particular, the notion of connectedness between a player's reasoning agent and his game agents is formally introduced to capture the assumption of sequential strategic stability, i.e. compliance with the initial strategy. According to this definition, an agent is either high-connected if he acts in line with the initial strategy or low-connected otherwise. Hence, beliefs about the connectedness of opponents' agents enters the belief space as an additional epistemic feature. Applying this framework, sufficient conditions for backward induction are obtained by explaining surprise information with low-connectedness of the deviating agent. Rather than revising the belief in an opponent's rationality, a supposedly irrational move of one of his agents at a preceding decision node is accommodated by belief revision on the high-connectedness of that agent, which, in turn, separates that supposedly irrational agent from the remaining agents of the respective opponent.

Various substantial interpretations of our framework become available. Interpreting sequence temporally, the three stages in a dynamic game reflect a player as existing over time: initially, a player assigns utilities to possible outcomes, subsequently chooses a strategy and at later points in time, he actually plays. In fact, players existing over time can be interpreted as multiple-selves and their agents as selves. Hence, theories of personal identity over time can be used to describe agent connectedness as intrapersonal connectedness in the multiple-self, with such features as degree of continuity of psychological features, memory or sympathy. Farther, the interpretation of our framework unveils strong assumptions implicit in the principle of epistemic independence which underlies any foundational argument for backward induction. Indeed, an observed surprise must never induce a belief revision on any intrapersonal connectedness of game agents at any later points in the game. Finally, our framework can be applied to the backward induction paradoxes by providing probabilities for future deviation of agents.

To illustrate our framework, consider the dynamic game of perfect information given by the extensive form given in Figure 1.

Such games are commonly solved by backward induction as follows. At *Alice*'s second decision node, her unique optimal choice is f. Given this choice of *Alice*, *Bob*'s unique optimal action at his decision node is d. Given the unique opti-

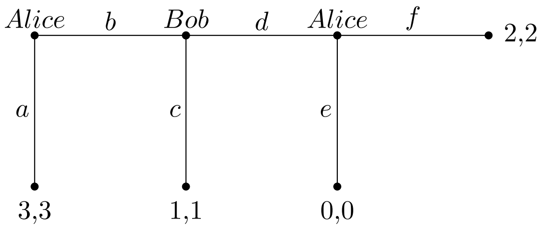

Fig. 1.

mal choices of *Alice* at her second decision node and *Bob* at his decision node, *Alice* picks a at her first decision node. The backward induction strategy profile (af, d) thus obtains. Note that *Bob* has to entertain the possibility of *Alice* having deviated from her backward inductive strategy when determining the choice for his decision node. Even though *Alice* could not have complied with her backward inductive strategy, *Bob* is assumed to think that *Alice* will nevertheless act in accordance with backward induction later on and hence to play the backward inductive move himself at his decision node. Accounting for the surprise that Alice has played b while still maintaining that she will play d is vital in making backward induction reasoning work. Usually, an assumption of epistemic independence is used to exclude any influence of such deviating behavior on expectations about *Alice*'s future behavior. In our account, surprise information is explained with low agent connectedness of the agent governing *Alice*'s first decision node. Since low agent connectedness can be interpreted by a multiple-self model of personal identity over time, player *Alice* can be understood as such a multiple-self. Accordingly, the deviating behavior of *Alice*'s first agent can be explained as exhibiting low intrapersonal connectedness. For instance, such low intrapersonal connectedness can occur due to a breakdown of psychological features, memory or sympathy between the self at *Alice*'s first decision node and her other selves. Such a more detailed description of dynamics allows us to more fully understand strategic interaction over time. In particular, the essence of backward induction can be elucidated.

We proceed as follows. Section 2 introduces our three-stage account of dynamic games, and defines a player as a set of agents. Section 3 describes players' reasoning by extending a type-based epistemic model of dynamic games with agent connectedness. Then, Section 4 gives sufficient conditions for backward induction in terms of agent connectedness. Section 5 discusses some interpretative issues of our framework, with particular emphasis on a multiple-self interpretation of a player. Finally, Section 6 offers concluding remarks.

2 Modeling Dynamic Games

Here, we make explicit the inherent sequential structure of dynamic games by a three-stage account that distinguishes between set-up, reasoning and play stages. Farther, a player is defined as a set of agents corresponding to these three stages, namely set-up agent, reasoning agent and game agents. Two notions of strategy are tailored to the reasoning and play stage, respectively. Our three-stage account unveils a sequential stability assumption implicit in the standard model of dynamic games.

Since Kuhn (1953), dynamic strategic interaction has commonly been modeled by the so-called extensive form which represents a game as a tree.

Definition 1. *An* extensive form structure with perfect information *is a tuple* $\Gamma = (X, Z, E, x_0, I, m, (u_i)_{i \in I})$*, where*

- X *is a finite set of non-terminal nodes, specifying decision nodes,*
- Z *is a finite set of terminal nodes, specifying the different situations in which the game may end,*
- E *is a finite set of directed edges* $(x, y) \in X \times (X \cup Z)$*, specifying the choices for the players, where* (x, y) *moves the game from* x *to* y*,*
- x_0 *is the unique root of the tree and called initial node,*
- I *is a finite set of players, where* $\mid I \mid > 1$*,*
- $m : X \setminus Z \to I$ *is the move function assigning to every non-terminal node the choosing player, where* X_i *denotes the set of all* $x \in X$ *such that* $m(x) = i$*,*
- $u_i : Z \to \mathbb{R}$ *is player* i*'s utility function assigning to every terminal node* $z \in Z$ *a utility* $u_i(z)$*.*

The extensive form can be interpreted as a set-up procedure for modeling a dynamic game.[1] First, the structure of the game has to be specified, i.e. all of its possible situations, outcomes at final situations, and rules are formalized by the sets X, Z and E, respectively. Then, a particular set of players determines the decision-makers in the game and the corresponding contingent situations where they act is given by the move function m. In a final step, each player has to consider all possible outcomes of the game and assign cardinal utilities to them in line with his preferences. To make explicit this procedural character inherent in the extensive form, we call the course of fixing the model *set-up stage* of a dynamic game. Once the game is fixed, the players can reason about it for decision-making purposes and thereafter the game is actually played.

A further basic ingredient when modeling dynamic games is the notion of strategy, which is considered the object of choice for the players. A strategy specifies an action for each contingency that might possibly arise for the respective player.

Definition 2. *Let* Γ *be an extensive form structure with perfect information,* $i \in I$ *some player and* $A_i(x_i) \subseteq E$ *the set of edges departing from* $x_i \in X_i$*. A* strategy *for* i *is a function* $s_i : X_i \to \bigcup_{x_i \in X_i} A_i(x_i)$ *such that* $s_i(x_i) \in A_i(x_i)$ *for all* $x_i \in X_i$*.*

[1] Note that we restrict attention to dynamic games with perfect information, i.e. games in which all players, whenever they have to choose, know exactly the choices made by their opponents until then.

According to the standard view, a strategy specifies an action for each contingency that might possibly arise for the respective player and hence can be interpreted as his disposition to act at each of his decision nodes. We call such a choice plan an *actual strategy* since it refers to the contingent actions of the players when actually playing the game. However, also before the game is played, players determine strategies based on their hypothetical reasoning. Such objects resulting from the players' reasoning and being fixed before play are called *initial strategies* and are formally defined in the next section. Note that actual strategies can differ from initial strategies.

After the set-up stage, a player reasons about his opponents as well as the fixed game, and decides on a complete contingent choice plan for the game as a result of this reasoning. We call this process the *reasoning stage* of a dynamic game and the player's ensuing hypothetical choice plan is his initial strategy. Note that although coming after the set-up stage, the reasoning stage is prior to the actual play of the game. The introduction of the reasoning stage thus explicitly separates hypothetical plans from actual choices.

After the set-up and reasoning stages the game is actually played and all contingent situations that may possibly arise in the game are represented in the extensive form by a player's set of decision nodes. We assume that each such node is governed by an agent of the player and call the actual playing phase of the dynamic game the *play stage*. With the game structure and initial strategy being fixed by the prior two stages, the play stage determines the strategy profile that is actually played as well as the corresponding outcome and utilities for the players. Hence, our account distinguishes between three stages of a dynamic game: the set-up stage, the reasoning stage and the play stage.

Farther, note how our three-stage view on dynamic games makes use of the notion of player. Accordingly, two distinguishable tasks are performed by a player before the play stage: utilities have to be assigned to outcomes in the set-up stage, followed by the choice of an initial strategy in the reasoning stage. During the play stage each of the decision nodes specifies a distinguishable task to be handled by one agent, respectively. In order to be able to discern the acting entities of the different stages, we understand the player as consisting of a *set-up agent*, a *reasoning agent* and *game agents*. Formally, a player is defined as the set of his agents.

Definition 3. *Let Γ be an extensive form structure with perfect information. A player $i \in I$ in Γ is defined as a set of agents $i = \{\alpha^i_s, \alpha^i_r, \alpha^i_1, \alpha^i_2, ..., \alpha^i_m\}$, where $\mid X_i \mid = m \in \mathbb{N}$, and α^i_s is called* set-up agent*, α^i_r is called* reasoning agent*, and all other agents α^i_j are called* game agents*, each corresponding to a unique decision node $x_i \in X_i$.*

The preceding definition of a player as a set of agents makes formally explicit the different tasks to be performed by a player in a dynamic game, related to the three different stages.

Note that our account of dynamic games makes transparent their sequential structure. Yet, a stability assumption lurks implicitly in the standard extensive form model. Despite the sequential character of dynamic games, no changes in the game's ingredients, utility assignments or choice prescriptions by the initial strategy is admitted during the dynamic interaction. In other words, any object fixed in the two pre-play stages, once determined, remains rigid until the end of the game. The basic game ingredients - decision nodes, strategy sets, possible outcomes, player sets and utility assignments - are assumed to remain stable throughout reasoning and play. In particular, invariant utilities reflect the assumption of stable preferences for all agents. Moreover, the deliberation of the reasoning agent of a player is supposed to apply to all game agents, who are all required to adhere to the initial strategy. Hence, any dynamics concerning the game structure as well as concerning the player are excluded by the standard extensive form model of dynamic games. While it may seem plausible to keep the game structure fixed given an underlying dynamic game to be modeled, the suspension of any dynamics concerning the player represents a rather strong assumption within the standard model. The introduction of three stages relevant to a dynamic game allows us to explicitly endorse or weaken the stability assumption with respect to deviation from pre-play reasoning.

The idea of understanding a player as a set of agents is now illustrated with the extensive form depicted in Figure 1. In addition to the game agents at their respective decision nodes, both players $Alice$ and Bob have two further agents that determine their utilities and strategies before play, corresponding to the set-up and reasoning stage, respectively. The two players can thus be formalized as sets $Alice = \{Alice_s, Alice_r, Alice_1, Alice_3\}$ and $Bob = \{Bob_s, Bob_r, Bob_2\}$. Actual choice of a player is then described by a strategy, each component of which is determined by the respective game agent in charge. For example, the actual strategy profile (be, d) signifies that $Alice_1$ chooses b at her first decision node, then Bob_2 picks d at his decision node and $Alice_3$ selects e at her second decision node. However, the initial strategies of the reasoning agents could be different. For instance, $Alice_r$ might have chosen bf prior to play. Note that in this example, a common index is used for both players to identify the position of their agents in the game tree and to reflect its sequential structure. More complicated game trees such as ones with parallel nodes governed by different agents of one player can then still be given some sequential order, relative to the structure of the game tree. Farther, game agents assigned to decision nodes that are excluded by actual play can be interpreted as inactive game agents. Also, it is possible to conceive of a player as having inactive agents at opponent decision nodes, and to hence interpret the player as a decision-maker over time with inactive agents at points in the game where no game agent acts for him. For instance, the set representing $Alice$ would then be amended with the inactive agent $Alice_2$, and the set representing Bob would be amended with the inactive agents Bob_1 and Bob_3, where the inactive agents correspond to decision nodes which are assigned to opponent game agents, respectively.

Our three-stage account of dynamic games proposed in this section makes explicit the sequential character of dynamic games and the stability assumptions already implicit in the standard extensive form model. A player is conceived of as a set of agents relative to the three sequential stages, making explicit that different agents of a player act in distinct sequential situations before and during play. Section 5 discusses various interpretations that can be used in the context of this account. For instance, the sequential structure of dynamic games can be understood as temporal, players can be perceived as persons or multiple-selves and initial strategies can be viewed as intentions of the respective players. The next section proposes an epistemic model for the reasoning stage of dynamic games and formalizes the notion of initial strategy.

3 Extending Type-based Interactive Epistemology

Interactive epistemology, also called epistemic game theory when applied to games, provides an abstract framework to formalize epistemic notions such as belief and knowledge. This rather recent field of research was initiated by Aumann (1976) and first adopted in the context of games by Tan and Werlang (1988). The fundamental problem addressed is the description of the players' choices in a given game relative to various epistemic assumptions. Epistemic game theory builds on the basic intuition that a player has to reason about the other players. Before choosing his strategy, he must form a belief about what his opponents will do. However, in order to so, he also needs to form a belief about what the others believe that their opponents will do. Similarly, any higher-order beliefs about his opponents are relevant to the player's choice. In order to formally represent players' reasoning about each other, an epistemic model is added to the analysis of a game. Here, we follow the type-based approach to epistemic game theory, according to which different epistemic states are encapsulated in the notion of type. More precisely, a set of types is assigned to every player, where each player's type induces a belief on the opponents' choices and types. Thus any higher-order belief can be derived from a given type. The notion of type was originally introduced by Harsanyi (1967-68) in the specific context of incomplete information but can actually be generalized to any interactive uncertainty. A recent survey of type-based interactive epistemology is provided by Siniscalchi (2008). Here, we extend the standard type-based epistemic model with the new notion of initial strategy.

Before our epistemic model can be defined, one further notion is needed. Letting S_j denote the set of all strategies of player j, a strategy $s_j \in S_j$ is said to avoid a given decision node $x \in X$, if there exists some decision node $x^* \in X$ on the unique path from the initial node x_0 to x, for which s_j assigns an off-path action. The set $S_j(x) \subseteq S_j$ then denotes all strategies of player j that do not avoid node x. An extended epistemic model for dynamic games can now be defined as follows.

Definition 4. *Let Γ be a finite extensive form structure with perfect information. An* extended epistemic model *of Γ is a tuple $\mathcal{M}^\Gamma = (T_i, \beta_i, \iota_i)_{i \in I}$, where T_i is a finite set of types for player i; $\beta_i : T_i \times (X_i \cup x_o) \rightarrow \Delta(\times_{j \in I \setminus \{i\}}(S_j \times T_j))$ assigns to every type $t_i \in T_i$ and decision node $x_i \in X_i$, a probability distribution* *on the set of opponents' strategy-type pairs, where* $\beta_i(t_i, x_i) \in \Delta(\times_{j \in i \setminus \{i\}}(S_j(x) \times T_j))$ *for all* $x \in X_i \cup \{x_0\}$; $\iota_i : T_i \to S_i$ *assigns to every type* $t_i \in T_i$ *an initial strategy.*

In the context of our three-stage account of dynamic games, the extended epistemic model concerns the reasoning stage. Thus, the deliberation of a player's reasoning agent is formalized by the extended epistemic model. In particular, the reasoning agent is disposed with conditional beliefs of any order at each of his decision nodes as well as the initial node, via the probability function β_i. Crucially, it is a distinguished feature of our epistemic model that a type does not only hold conditional beliefs about the opponents' *actual* strategies, but also about the opponents' *initial* strategies. Note that the conditional beliefs of the reasoning agent refer to hypothetical epistemic states of the respective game agents. Hence, while types and their induced conditional belief hierarchies model the deliberation process of the reasoning agent, the novel ingredient of initial strategy is interpreted as the outcome of the player's reasoning. Farther, note that the conditional beliefs of a player i at a given node $x \in X \cup \{x_0\}$ only assign positive probability to opponents' strategy choices that do not avoid x. This seems reasonable since otherwise a player would exhibit contradictory beliefs: although knowing to be at decision node x, he believes that at least one opponent has chosen a strategy avoiding x and thus excluding it to be reached.

The initial strategy is fixed in the reasoning stage before the play stage, in which the game is actually played. Choices by the game agents might differ from the ones prescribed by the reasoning agent's initial strategy. Since a player is conceived of as a set of agents by Definition 3, such a behavioral deviation from the initial strategy raises the problem of connectedness between a player's game agents and his reasoning agent. On the basis of the initial strategy function, we now formally introduce connectedness into our extended epistemic model for dynamic games.

Definition 5. *Let* $\mathcal{M}^{\Gamma}$ *be an extended epistemic model of an extensive form structure* Γ *with perfect information. Further, let* $\iota_i^{t_i}(x_i)$ *denote the action that the initial strategy of type* $t_i \in T_i$ *designates for game agent* $\alpha^i_{x_i}$ *at* $x_i \in X_i$*, let* s_i^{α} *denote the strategy of* i *that is actually played and let* $s_i^{\alpha}(x_i)$ *denote the actual choice of game agent* $\alpha^i_{x_i}$ *at* x_i*. The* connectedness $c_i(\alpha^i_{x_i}, s_i^{\alpha} \mid t_i)$ *of game agent* $\alpha^i_{x_i}$ *is defined as*

$$c_i(\alpha^i_{x_i}, s_i^{\alpha} \mid t_i) = \begin{cases} \textit{high} & \textit{if } \iota_i^{t_i}(x_i) = s_i^{\alpha}(x_i), \\ \textit{low} & \textit{otherwise.} \end{cases}$$

In the above definition, the actual strategy played refers to the actual choices of the respective player's game agents at the decision nodes they govern. Initial and actual strategy are then compared. A game agent is said to be *high-connected* if he acts in compliance with the initial strategy and *low-connected* otherwise. Connectedness hence both separates and relates sequential parts of the player at contingent points of the game. Note that the connectedness function expresses a behavioral notion as its values are determined by the actual choices of the game agents, relative to the initial strategy of the reasoning agent. In this context, the reasoning agent can be seen as the central representative of the player. This is plausible as the reasoning agent initially chooses a complete strategy that is intended to apply throughout the game, whereas the game agents only act locally. Also, stability of the initial strategy and hence equivalence to the actual strategy is implicitly assumed in the standard extensive form model. In the sequel, we therefore refer to reasoning agent and player interchangeably.

Moreover, note that the above definition can be generalized to provide a more realistic interpretation of a player in a dynamic game as a person. For instance, departing from a purely behavioral notion of connectedness, it is possible to furnish additional interpretation, such as underlying connectedness of psychological features, memory and sympathy. In this case, more general formal definitions of the connectedness function are available. For example, connectedness could be defined to be any subset of the real numbers or any countable or even uncountable abstract set. Also, rather than focusing on the relation between the reasoning agent and the game agents of a given player, connectedness between any pair of agents can be considered. Such interpretations are addressed in Section 5.

In type-based epistemic models, the objects of beliefs are events. Intuitively, an event states a property concerning the model's uncertainty space. Within the context of games, examples of events are "*Alice* plays strategy bf", "*Bob* is rational" and "*Bob* believes at the initial node that *Alice*'s agents are high-connected". Formally, events are simply sets of types. More precisely, a set $E \subseteq \bigcup_{i \in I} T_i$ of types is called event. The belief of some player i at some node $x \in X_i \cup \{x_0\}$ in some event E can then be modeled by projecting $\beta_i(t_i, x_i)$ on T_{-i}, denoted as $\beta_i(t_i, x_i \mid T_{-i})$. Similarly, player i's belief at some node $x \in X_i \cup \{x_0\}$ on the type of player $j \in I \setminus \{i\}$ can be obtained by projecting $\beta_i(t_i, x_i)$ on T_j, denoted as $\beta_i(t_i, x_i \mid T_j)$. Moreover, player i's belief at some node $x \in X_i \cup \{x_0\}$ on player j's strategy-type pair can be extracted by projecting $\beta_i(t_i, x_i)$ on $S_j \times T_j$, denoted as $\beta_i(t_i, x_i \mid S_j \times T_j)$. Note that beliefs are events, too and that indeed any higher-order belief can be represented in a type-based epistemic model. Given some event, a player's type specifies conditional belief hierarchies at each of his decision nodes. Epistemic states are thus local and concern the respective node-governing agent of the player. Yet they are hypothetical in the sense of belonging to the reasoning agent when deliberating before play about what his game agents would know were their respective nodes be reached.

For the purpose of formalizing rationality in our framework, let $u_i(\iota_i(t_i), \beta_i(t_i) \mid x_i)$ denote player i's expected utility starting at node x_i of playing the relevant part of strategy $\iota_i(t_i)$ given his belief at x_i about the opponents' strategies.

Definition 6. *Let $\mathcal{M}^{\Gamma}$ be an epistemic model of an extensive form structure Γ with perfect information and $i \in I$ some player. A type $t_i \in T_i$ is* rational *if $t_i \in R_i = \{t_i \in T_i : u_i(\iota_i(t_i), \beta_i(t_i) \mid x_i) \geq u_i(s_i, \beta_i(t_i) \mid x_i)$ for all $s_i \in S_i$ and for all $x_i \in X_i\}$.*

Accordingly, a type of a player is rational if his initial strategy maximizes his expected utility at every decision node in the game. Rationality is hence understood as a notion relative to the result of a player's reasoning, since it is precisely the outcome of his reasoning that reflects his attitude towards the interactive situation he is involved in. Note that our notion of rationality proposed here is weaker than the standard one, since the latter requires actual choice to be optimal throughout the tree while the former only concerns initial choice. A player can thus be rational in our sense while still actually acting irrationally in the standard sense. As an illustration of this observation, consider the game given in Figure 1. Suppose *Alice* believes that *Bob* chooses d and her reasoning agent $Alice_r$ hence picks the rational initial strategy af. Nevertheless, $Alice_1$ can still choose b at her decision node, hence acting irrationally in the standard sense.

As has already been pointed out above, within our extended epistemic framework a player can be perceived as the reasoning agent and his object of choice, the initial strategy, can be perceived as the result of the reasoning process, for instance, as his intention or plan of action for the game. A player's game agents then actually choose actions at the respective decision nodes, either in line with the initial strategy of the reasoning agent, or differently. Specific patterns of relationship between a player or his reasoning agent and his game agents can be formalized. We call a player $i \in I$ *high-connected* if all of his game agents are highly connected, i.e. $c_i(\alpha^i_{x_i}, s^\alpha_i \mid t_i) = \mathit{high}$ for all $\alpha^i_{x_i} \in i$. In other words, the game agents of a high-connected player actually choose in complete accordance with his proposed initial strategy. However, it is possible that only some game agents are highly connected, while others are not. For instance, only game agents succeeding some particular node might be high-connected. Crucially, belief in different patterns of high-connected game agents can be defined in our model. For instance, the following condition requires a player to believe in the high-connectedness of an opponent at all future nodes.

Definition 7. *Let* $\mathcal{M}^{\Gamma}$ *be an epistemic model of an extensive form structure* Γ *with perfect information,* $i \in I$ *be some player, and* $x \in X_i \cup \{x_0\}$ *some node. A type* $t_i \in T_i$ believes in j's future-high-connectedness at node x *if* $t_i \in BH_{i,j}(x) = \{t_i \in T_i : supp(\beta_i(t_i, x \mid S_j \times T_j)) \subseteq \{(s_j, t_j) \in S_j \times T_j : s_j(x_j) = \iota_j^{t_j}(x_j)$ *for all* $x_j \in X_j$ *succeeding* $x\}\}$.

This definition relates a player's belief on what an opponent is actually playing with the belief about his initial strategy. More generally, our model is also capable of distinguishing between actual and initial strategy in the reasoning of players about their respective opponents.

The preceding definitions introduce connectedness of game agents to their player, conceived of as the reasoning agent, into our extended epistemic framework, which in turn can be used to understand reasoning in strategic interaction over time. In a first such step, these notions are used in the next section to shed light on backward induction reasoning in dynamic games with perfect information.

4 Sufficient Conditions for Backward Induction

Backward induction constitutes the standard reasoning method in dynamic games with perfect information: at each decision node, optimal behavior is determined by assuming the optimality of choices at all succeeding nodes. Before formally defining backward induction we restrict attention to generic games with perfect information.

Definition 8. *An extensive form structure Γ with perfect information is called* generic *if for every player $i \in I$, for every decision node $x_i \in X_i$, for every two actions $a_i, a_i' \in A_i(x_i)$, every two terminal nodes $z \in Z$ such that z follows a_i and z' follows a_i', it holds that $u_i(z) \neq u_i(z')$.*

Accordingly, any two different choices at a given decision node will always lead to two distinct utilities for the respective player. It is common to assume genericity when searching for epistemic characterizations of backward induction. Since genericity implies uniqueness of the backward inductive strategy profile, no ambiguity arises in determining the actions in line with backward induction at each node in the tree. This restriction is not severe, since the aim is to unveil the epistemic states portraying the way of thinking characteristic of backward inductive reasoning. Genericity avoids the introduction of somewhat arbitrary criteria for ties that would divert from the essential properties of the players' reasoning required for backward induction to obtain.

Farther note that backward induction can only be defined for finite games, as possible end points of the game are required for the backward inductive process to begin. Finiteness is already implicit in our definition of the extensive form.

In order to facilitate the formal expression of backward induction, the decision nodes are classified according to their maximal distance from an end point i.e. a terminal node of the game, independent from any closer terminal nodes.

Definition 9. *Let Γ be an extensive form structure with perfect information and $x \in \bigcup_{i \in I} X_i$ some decision node. Decision node x is called* ultimate *if x is only immediately succeeded by terminal nodes; decision node x is called* pre-ultimate *if x is only immediately succeeded by ultimate decision nodes or by ultimate decision nodes and terminal nodes; decision node x is called* pre-pre-ultimate *if x is only immediately succeeded by pre-ultimate decision nodes or by pre-ultimate decision nodes and ultimate decision nodes or by pre-ultimate decision nodes and terminal nodes or by pre-ultimate decision nodes and ultimate decision nodes and terminal nodes; etc. Decision node x is called* initial node *if $x = x_0$.*

It is now possible to define backward induction for generic finite dynamic games of perfect information as follows.

Definition 10. *Let Γ be a generic extensive form structure with perfect information, $i \in I$ some player and $x_i \in X_i$ some decision node of i. The unique backward inductive choice $b_i(x_i) \in A_i(x_i)$ at x_i is determined as follows: if x_i is an ultimate node, then $b_i(x_i)$ is the unique action that maximizes i's utility* *at x_i, and if x_i is pre-ultimate node, then $b_i(x_i)$ is the unique action at x_i that maximizes i's utility at x_i given backward inductive actions at all decision nodes succeeding x_i, etc. Player i's unique* backward inductive strategy $b_i \in S_i$ *assigns to each of i's decision node $x_i \in X_i$ the respective unique backward inductive action $b_i(x_i) \in A_i(x_i)$.*

It is natural for epistemic game theory to search for epistemic requirements that induce the players to choose their backward inductive strategies. Indeed, various different sufficient conditions for backward induction have been proposed in the literature, which are reviewed, unified and compared by Perea (2007). Farther, note that the emphasis lies on what requirements are needed for a player to actually *choose* his backward inductive strategy and hence to make transparent the complete reasoning underlying backward induction. The genuinely different question of what epistemic conditions are needed to get the backward inductive *outcome* is addressed in, for instance, Battigalli and Siniscalchi (2002) and Brandenburger et al (2008). Here, we give an epistemic characterization of the backward inductive strategy profile in terms of connectedness.

Some more epistemic concepts need to be introduced before formal conditions for backward induction can be stated.

Definition 11. *Let $\mathcal{M}^{\Gamma}$ be an epistemic model of an extensive form structure Γ with perfect information, and $i \in I$ some player. A type $t_i \in T_i$* structurally believes in his opponents' rationality *if $t_i \in SBR_i = \{t_i \in T_i : supp(\beta_i(t_i, x \mid T_j)) \subseteq R_j$, for all $x \in X_i \cup \{x_0\}$, for all $j \in I \setminus \{i\}\}$.*

According to the above definition, at the beginning of the game as well as at any of his decision nodes, a player believes that all of his opponents are rational i.e. choose a rational initial strategy.

Iterating structural belief in rationality gives the nested epistemic notion of common structural belief in rationality.

Definition 12. *Let $\mathcal{M}^{\Gamma}$ be an epistemic model of an extensive form structure Γ with perfect information and $i \in I$ some player. A type $t_i \in T_i$ expresses* common structural belief in rationality *if $t_i \in CSBR_i = \{t_i \in T_i : t_i \in SBR_i^k$ for all $k \geq 1\}$, where $SBR_i^1 = SBR_i$, and $SBR_i^{k+1} = \{t_i \in T_i : supp(\beta_i(t_i, x \mid T_j)) \subseteq SBR_j^k$, for all $x \in X_i \cup \{x_0\}$, for all $j \in I \setminus \{i\}\}$, for all $k \geq 1$.*

Intuitively, the event of player i satisfying common structural belief in rationality describes the situation in which i initially as well as at each of his decision nodes, believes that his opponents initially choose rationally, i.e. optimal everywhere in the game tree, initially as well as at each of his decision nodes, believes that his opponents initially as well as at each of their decision nodes believe that their opponents initially choose rationally i.e. optimal everywhere in the game tree, etc. In other words, player i always believes that his opponents choose optimal initial strategies, always believes that every opponent always believes that every other player always chooses an optimal initial strategy, etc. Observe that due to our weaker notion of rationality in Definition 6, it is always possible to define common structural belief in rationality in our epistemic model, contrary to impossibility results, such as by Reny (1992) and (1993), concerning epistemic models with standard rationality. While it is usually not distinguished between initial and actual choice, common structural belief in rationality cannot be generally defined in standard epistemic structures. However, our model is capable of admitting that a player believes that an opponent initially chooses rationally, while at the same time entertaining the belief that the same opponent will actually choose irrationally at some points in the game and thus not carry out the rational strategy of his respective reasoning agent. In our model, a player can reason about both the reasoning as well as the actual play of his opponents. In this context note that Perea (2008) provides an epistemic model in which common structural belief in standard rationality is generally made possible by allowing a player to revise his beliefs about his opponents' utilities during the game, while assuming the respective player's utilities to be constant. As an illustration of the permanent feasibility of common structural belief in rationality in our framework, consider the dynamic game given in Figure 1. Suppose satisfying common structural belief in rationality, *Bob* believes at the beginning of the game as well as at his decision node that *Alice* initially rationally chooses strategy af. It is then possible that *Bob* believes at his decision node that *Alice* actually chooses a strategy different from af, i.e. that game agent $Alice_1$ has picked b and game agent $Alice_3$, for instance, will pick e, while still maintaining his belief in $Alice_r$'s rational choice of the initial strategy af.

Farther, note that games are epistemically investigated from a perspective which is completely that of a single player. Even nested belief notions are defined from the viewpoint of a specific player. Understanding interactive epistemology as a theory of reasoning prior to choice, this stance seems natural, since any reasoning process takes place entirely within the reasoning individual, represented here by the reasoning agent.

Connectedness is now used to define the nested epistemic notion of forward belief in future-high-connectedness.

Definition 13. *Let $\mathcal{M}^{\Gamma}$ be an epistemic model of an extensive form structure Γ with perfect information and $i \in I$ some player. A type $t_i \in T_i$ expresses* forward belief in future-high-connectedness *if $t_i \in FBH_i = \{t_i \in T_i : t_i \in BH_i^k(x)$, for all $k \geq 1$, for all $x \in X_i \cup \{x_0\}\}$, where $BH_i^1(x) = \{t_i \in T_i : t_i \in BH_{i,j}$ for all $j \in I \setminus \{i\}\}$, and $BH_i^{k+1}(x) = \{t_i \in T_i : supp(\beta_i(t_i, x \mid T_j)) \subseteq BH_j^k(x_j)$, for all $j \in I \setminus \{i\}$, for all $x_j \in X_j$ such that x_j follows $x\}$, for all $x \in X_i \cup \{x_0\}$, and for all $k \geq 1$.*

According to forward belief in future-high-connectedness, a player always believes that his opponents' agents are highly connected at all succeeding nodes, that his opponents believe at all succeeding nodes that their opponents-agents are highly connected at all respectively succeeding nodes, etc. Observe that this epistemic condition implies that at any possible situation in the game, the player believes that any opponent agent at a succeeding decision node is highly connected and hence acts in accordance with the respective initial strategy of his player.

Farther, note generally that requiring forward belief in some event E is a considerably weaker epistemic condition than common structural belief in E. Accordingly, a theorem only requiring forward belief in some particular event and not common structural belief is strengthened. To see that common structural belief in E is stronger than forward belief in E, consider a decision node x_i succeeding some node x_j. According to the former epistemic condition i believes at x_i that opponent j believes E at x_j, while the latter concept of forward belief in E does not put any restrictions on what i believes at x_i what j believes at any preceding decision node, in particular not whether j believes E at x_j. Intuitively, the strength of common structural belief derives from the fact that it concerns any decision node, including respectively preceding ones, relative to a given decision node. In contrast, forward belief concerns only succeeding decision nodes, given a particular decision node.

It is now possible to formulate epistemic conditions for backward induction in terms of connectedness.

Theorem 1. *Let $\mathcal{M}^{\Gamma}$ be an epistemic model of a generic extensive form structure Γ with perfect information and $i \in I$ some player. If $t_i \in T_i$ such that $t_i \in R_i \cap CSBR_i \cap FBH_i$, then $\iota(t_i) = b_i$.*

Proof. Suppose that $t_i \in T_i$ such that $t_i \in R_i \cap CSBR_i \cap FBH_i$. We show that t_i initially assigns the backward inductive action to each decision node $x_i \in X_i$, i.e. $\iota(t_i) = b_i$. Consider a decision node $x_i \in X_i$ of player i. Suppose that x_i is an ultimate decision node. Then, by rationality of t_i, $\iota_i^{t_i}(x_i) = b_i(x_i)$. Suppose that x_i is a pre-ultimate decision node. Since $t_i \in FBH_i$, type t_i believes at x_i that every opponent game agent $\alpha^j_{x_j}$ is high-connected to his respective player j and hence chooses according to j's initial strategy at every ultimate decision node x_j succeeding x_i. As $t_i \in CSBR_i$, he also believes at x_i in j's rationality i.e. that j's initial strategy is rational. Hence, t_i believes that every high-connected opponent game agent $\alpha^j_{x_j}$ does indeed choose rationally at every x_j succeeding x_i, and thus picks the unique backward inductive action $b_j(x_j)$ there. Therefore, the unique optimal action for i at x_i is the backward inductive one and rationality of t_i ensures that $\iota_i^{t_i}(x_i) = b_i(x_i)$. Now suppose that x_i is a pre-pre-ultimate decision node. Since $t_i \in FBH_i$, type t_i believes at x_i that every opponent game agent $\alpha^j_{x_j}$ is high-connected to his respective player j and hence chooses according to j's initial strategy at every decision node x_j succeeding x_i. Note that every opponent decision node x_j succeeding x_i is either pre-ultimate or ultimate. Suppose that x_j is ultimate. As $t_i \in CSBR_i \cap FBH_i$, type t_i believes at x_i in j's rationality i.e. that j initially chooses rationally, as well as that every high-connected opponent game agent $\alpha^j_{x_j}$ does indeed choose rationally at every ultimate decision node x_j, and thus picks the unique backward inductive action $b_j(x_j)$ there. Suppose that x_j is pre-ultimate. Since $t_i \in FBH_i$, type t_i believes at x_i that at any immediately succeeding opponent decision node x_j, the respective opponent j believes that his opponents' game agents are high-connected, and thus act in accordance with their respective player's initial strategy, at all succeeding ultimate decision nodes. Also, by $t_i \in CSBR_i$, type t_i believes at x_i

that his opponents believe at all succeeding nodes in their opponents' rationality i.e. that their opponents have initially chosen rationally. Hence, t_i believes at x_i that at any immediately succeeding opponent decision node x_j, the respective opponent j believes that his opponents' high-connected game agents play rationally at every ultimate decision node succeeding x_j. Moreover, by $t_i \in CSBR_i$, type t_i also believes at x_i in j's rationality, i.e. in a rational initial strategy choice of j. Since $t_i \in FBH_i$, it then follows that he believes at x_i that every high-connected game agent $\alpha^j_{x_j}$ does indeed choose rationally at the respective pre-ultimate decision node x_j. But as t_i believes at x_i that j believes at x_j that j's opponents choose the backward inductive actions at all ultimate decision nodes succeeding x_j, in fact t_i believes at x_i that $\alpha^j_{x_j}$ picks his unique backward inductive action $b_j(x_j)$ at x_j. Therefore, since t_i believes at x_i that at any succeeding decision node the respective opponent game agent chooses the backward inductive action, the unique optimal action for i himself at x_i is the backward inductive one and $\iota_i^{t_i}(x_i) = b_i(x_i)$ obtains by rationality of t_i. By induction, it follows that at any $x_i \in X_i$, type t_i believes that his opponent game agents choose the unique backward inductive action at any x_j succeeding x_i, and hence, being rational, t_i initially assigns the unique backward inductive choice to each of his decision nodes, i.e. $\iota_i(t_i) = b_i$. □

In our enriched epistemic framework, the preceding theorem provides a foundation for backward induction in terms of connectedness. Intuitively, common structural belief in rationality ensures that the respective player always believes that his opponents initially play rationally i.e. their unique backward inductive strategies, while at the same time he also always believes that his opponents' future game agents actually choose accordingly, by forward belief in future-high-connectedness. Then, i initially chooses his unique backward inductive strategy. In fact, any surprise information that might arise during play is explained by low-connectedness of the deviating game agent, maintaining the belief in future-high-connectedness of all succeeding game agents.

When reasoning about his opponents in the reasoning stage, a player's reasoning agent contemplates both about his opponents' reasoning as well as their actual choices. In fact, it is his conclusion on his opponents' actual choices that finally matters for the decision problem of the player's reasoning agent on the basis of which he then chooses an initial strategy. Conceptually, a type furnished by an epistemic model captures the complete reasoning of the respective player. Indeed, the epistemic states and the reasoning of a player coincide. During the play stage, agents then pick actual choices according to which the dynamic game unfolds. Importantly, actual decisions need not to be in accordance with the underlying reasoning. For instance, a player might change with regards to properties of his mental set-up such as psychological, emotional or memory features. These interpretative issues will be addressed in Section 5.

An epistemic model only prescribes a player's beliefs and intentions, i.e encompasses his reasoning, but it does not prescribe actual choices. Here, our framework precisely captures this basic idea of an epistemic model by distinguishing between initial and actual strategy choice by a reasoning agent and a set of game agents, respectively, and explicitly endorses the possibility of change in a player's decisions. A player's decision furnished by our epistemic model is accurately his initial strategy. Hence, the epistemic foundation for backward induction provided by Theorem 1 does concern a player's initial and not his actual strategy. However, our framework also permits the formulation of sufficient conditions for backward induction in terms of actual choice as follows.

Corollary 1. *Let $\mathcal{M}^{\Gamma}$ be an epistemic model of a generic extensive form structure Γ with perfect information. If $c_i(\alpha^i_{x_i}, s^{\alpha}_i \mid t_i) = high$, for all $x_i \in X_i$, and for all $i \in I$, as well as $t_i \in T_i$ such that $t_i \in R_i \cap CSBR_i \cap FBH_i$ for all $i \in I$, then $s^{\alpha} = b$.*

Proof. Consider $i \in I$ and suppose that $t_i \in T_i$ such that $t_i \in R_i \cap CSBR_i \cap FBH_i$. It follows from Theorem 1 that $\iota^{t_i}_i(x_i) = b_i(x_i)$ for all $x_i \in X_i$. Since $c_i(\alpha^i_{x_i}, s^{\alpha}_i \mid t_i) = \text{high}$, for all $x_i \in X_i$, each high-connected game agent of player i will indeed choose the backward inductive action $s^{\alpha}_i(x_i) = b_i(x_i)$ at any $x_i \in X_i$, respectively. Therefore, i's actual backward inductive strategy choice $s^{\alpha}_i = b_i$ obtains. □

Accordingly, the backward inductive strategy profile will be played if each player's reasoning agent is rational, expresses common structural belief in rationality as well as forward belief in future-high-connectedness, and each player's game agents are highly connected i.e. actually do carry out their reasoning agent's initial strategy. Again note that actual choice is a property of game agents not of the reasoning agent i.e. the type.

As an illustration of this epistemic foundation of backward induction, consider *Bob*'s reasoning in the dynamic game given in Figure 1. In order to choose his initial strategy in the reasoning stage, Bob_r hypothetically considers his game agent Bob_2. By forward belief in future-high-connectedness, Bob_r believes at Bob_2 that $Alice_3$ will be high-connected and thus play in line with the initial strategy of her reasoning agent $Alice_r$. Since, by common structural belief in rationality, he believes $Alice_r$ to choose rationally, Bob_r believes at Bob_2 that $Alice_r$ initially chooses f at her final decision node. Therefore, Bob_r believes at Bob_2 that the high-connected $Alice_3$ complies with the initial rational strategy and thus picks f. Hence, Bob_r initially chooses his rational strategy d, as well as actually in case of Bob_2 being high-connected.

Farther observe that our theorem makes explicit a strong principle of epistemic independence needed for backward induction: the observation of a deviating opponent game agent has no influence whatsoever on a player's beliefs concerning any game agents at succeeding decision nodes, who are still believed to be highly connected each. Also note that only requiring forward belief in future-high-connectedness instead of the stronger condition of common structural belief in future-high-connectedness strengthens our epistemic characterization of backward induction. The epistemic foundation for backward induction in terms of connectedness provided here is interpreted and discussed in Section 5.

5 Discussion

5.1 Dynamics

Our extended epistemic framework, which understands players as sets of agents and models their connectedness, is capable of shedding light on the *dynamic* character of dynamic games.

In a general sense, our framework displays the complete sequential structure underlying the standard extensive form model of dynamic games. According to our framework, a dynamic game has at least three distinguishable stages: the set-up stage, the reasoning stage and the play stage. It is thus made explicit that different agents of a player find themselves in distinct, sequential situations, such as utility assignments before play, reasoning before play and then play at different decision nodes. Moreover, explicating the sequential structure of dynamic games within our framework reveals stability assumptions implicit in the standard extensive form model. The ingredients of the game, including the fact that utilities are determined prior to reasoning and actual play as well as that pre-play strategy choice resulting from reasoning, are supposed to remain invariant during the whole dynamic strategic interaction. Concerning utilities, the assumption of stable preferences of all agents throughout reasoning and play is made explicit by the fact that they respond to the same utility function. Concerning reasoning, our model clarifies that a game agent is presumed at his decision node to comply with his player's instructions i.e. to act in line with the respective initial strategy.

The assumed stability of dynamic games implicit in the standard extensive form model can be argued to be in tension with its inherent sequential nature. While the latter suggests the possibility of change in dynamic games, the former does not offer enough structure to account for any such changes. This problematic aspect of sequential stability is made explicit and can be relaxed in our framework. In particular, the notion of high-connectedness can be formally introduced which captures the sequential stability of game agents. Intuitively, high-connectedness captures the idea that game agents make choices according to the pre-play instructions of their reasoning agents. Also, Theorems 1 and 2 relate high-connectedness to backward induction reasoning. Note that high-connectedness is a purely behavioral assumption that can be dropped locally, in order, for instance, to account for surprise information in backward induction reasoning. More generally, connectedness can be used to formulate hypothetical reasoning patterns related to the sequential stability of a player and his game agents, which in turn can be applied to epistemic characterizations of game-theoretic solution concepts.

Moreover, by clarifying the sequential character of dynamic games, unveiling stability assumptions and modeling reasoning about connectedness of agents, our framework provides foundations for a realistic interpretation of dynamic games as formal representations of strategic interaction over time. More specifically, two interpretative directions can be taken. Firstly, the very sequential structure of dynamic games as rendered transparent in our framework can be interpreted as temporal. Secondly, the player which is defined as a set of agents with specific tasks in our framework can be interpreted as a person. In particular, interpretations of players in dynamic games can be linked to multiple-self models of personal identity over time by understanding players as multiple-selves and agents as selves. The idea of decision-makers as multiple-selves can thus be addressed in our account of dynamic games. Indeed, the subsequent subsection interprets dynamic games from a multiple-selves point of view.

5.2 Multiple-Self

The conception of player as a set of connected agents can naturally be linked to the notion of connected selves in a multiple-self. Indeed, this idea of understanding agents of players as the different selves of multiple-selves has been employed in the context of extensive form models with imperfect information, such as that of Piccione and Rubinstein (1997). Also, the idea that a player consists of different acting selves appears in Selten (1975) and Halpern (2001) within the context of the agent normal form. Yet such appearances of the multiple-self concept in game theory lack philosophical foundations. Here, we propose to interpret the notion of player as a multiple-self using theories of personal identity over time, in order to give specific meaning to change of players over time and to the reasoning of players about possible or observed changes of their opponents.

The notion of the multiple-self is studied in the context of theories of personal identity over time, for instance Noonan (2003) or Raymond and Barresi (2003). Such theories investigate how a person both persists and changes over time. The seemingly contradictory nature of sameness and change in persons generates two main concerns in theories of personal identity over time. Firstly, emphasis is either given to sameness or to change of persons over time. At the extreme ends of the spectrum, theories focus either exclusively on sameness, such as the idea that persons are constituted by the soul which is assumed to be stable over time, or on difference over time, such as Hume's (1739) idea that there is great variation between different time slices within persons. Secondly, various criteria have been adopted to describe the substantive nature of personal identity over time. Examples for such criteria include different psychological features, the body, the brain, memory, emotions and consciousness.

Multiple-self theories, such as those of Parfit (1984), Elster (1986), and Ainslie (1992), understand persons as distinct yet interconnected selves. In these accounts, selves are capable of reasoning and acting, and are interconnected with each other to form a multiple-self. A multiple-self model of personal identity over time consists of three elements: a set of selves, a notion of intrapersonal connectedness between these selves and an interpretation of connectedness through a criterion of personal identity over time. These models can be related to our extended epistemic framework by interpreting the agents as selves and the player as a multiple-self. The purely behavioral notion of connectedness as measuring compliance or deviation of a game agent with or from the initial strategy of the reasoning agent can then be explained by *underlying connectedness* which describes the degree of continuity of psychological features, memory and sympathy. Such substantive interpretations of intrapersonal connectedness render the description of decision-makers in dynamic games more realistic. Adopting such a multiple-self model of personal identity over time, the three substantive interpretations of underlying connectedness are now considered and linked to our extended epistemic model and sufficient conditions for backward induction.

Psychological connectedness, mainly due to Parfit (1984), measures the degree of similarity between psychological traits of different selves, such as preferences. Accordingly, the temporal self is depicted as an agent acting on the basis of his preferences. In the context of game theory, psychological connectedness permits the interpretation of players as consisting of agents who govern decision nodes. These agents have the capacity to act according to their possibly different preferences. Furthermore, a supposedly irrational move at some decision node can be interpreted as resulting from different preferences of the respective agent. As an illustration consider the game given in Figure 1 and suppose that *Bob* believes $Alice_1$ to rationally play a. Upon observing a surprising move b by $Alice_1$, he can make sense of the low behavioral connectedness of $Alice_1$ as follows: Bob adopts the belief that $Alice_1$ has exhibited deviating preferences from her player. In particular, it is natural to depict a breakdown in psychological connectedness between $Alice_1$ and $Alice_s$. However, it could also be the case that $Alice_1$ has re-evaluated outcomes at later terminal nodes rendering her preferences different from $Alice_3$. Note that a particular contemplation about what precisely has prompted the preference change or about what precisely it consists in, is not formally needed to obtain our sufficient conditions for backward induction in terms of connectedness. Yet, in order to further describe backward induction reasoning, it is possible to provide such more realistic interpretations when viewing a player as a multiple-self. In order to obtain backward induction, an agent who observes a supposedly irrational move can hence maintain belief in the rationality of the respective opponent, as well as belief in future-high-connectedness by revising his belief in the high psychological connectedness of the deviating game agent. Such belief revision only commits the reasoner to believing there to have been a relevant preference change such as to prompt a local re-evaluation of the payoffs which, in turn, has led to a local deviation from the initial strategy. Note that similarly, Perea (2008) provides an epistemic model for dynamic games in which the possibility of belief change about opponents' utilities during the game is explicitly endorsed, modeled and sufficient conditions for backward induction are derived.

Sympathetic connectedness, such as proposed by Schechtmann (2001), measures the degree to which temporal selves can sympathize with each other. Such a sympathetic access expresses the strength of emotional bond between selves, supervening on physical and psychological features. In the context of game theory, a supposedly irrational move at an earlier decision node can plausibly be interpreted as a local breakdown in the opponent's sympathetic connectedness. By ascribing a low sympathetic connectedness to some self of another person, it is reasonable to still grant full rationality and reasoning capacity to the remaining game agents and the reasoning agent of that person. Similar interpretations for the dynamic game given in Figure 1 as proposed for psychological connectedness are hence available.

Memory connectedness, as originally proposed by Locke (1694) and further developed by Shoemaker and Swinburne (1984), measures the degree to which a self remembers having had an experience at an earlier time and thus expresses the extent of access to experiences of earlier selves. In the context of game theory, a supposedly irrational move at an earlier decision node can be interpreted as a breakdown of memory between the deviating agent and his player, in particular, that the deviating agent has forgotten the initial strategy.[2] related to imperfect recall. By assigning a low access to earlier experiences of the deviating agent, a reasoner can revise his belief in the stability of memory of an opponent, excluding the agent from the opponent's agents that share memories while still maintaining belief in rationality. Similar interpretations for the game given in Figure 1 as proposed for psychological and sympathetic connectedness are thus available.

The different interpretations of intrapersonal connectedness in the multiple-self allow both a more fine-grained discussion of reasoning about opponents as well as a more specific interpretation of how observed surprise moves at preceding decision nodes can be explained. Note that by adopting and combining further criteria of personal identity over time, as reviewed by Noonan (2003) and Raymond and Barresi (2003), various realistic accounts and interpretations of reasoning in dynamic games are possible in our framework.

5.3 Epistemic Independence

The extended epistemic framework proposed here allows us to characterize backward induction in terms of connectedness. This notion of connectedness measures the extent to which game agents are sequentially stable relative to the reasoning agent of the player. In Theorem 1, backward induction is assured by explaining surprise information in terms of low-connectedness of the deviating game agent.

Connectedness reflects the fundamental principle underlying any foundation of backward induction. This so-called principle of epistemic independence, which is conceptually discussed by Stalnaker (1998), requires that a player treats any information obtained during the game, such as observed opponents' moves, as irrelevant to his beliefs about opponents' behavior at later points in the game. This property is at work in our theorem: the observation of a surprising move of an opponent's game agent does not affect a player's beliefs on the behavior of the respective opponent's future game agents, but rather the concerned game agent is concluded to be low-connected. In other words, his comportment is regarded as isolated and irrelevant to future behavior of the represented player.

More specifically, forward belief in future-high-connectedness yields the condition of epistemic independence that is implicit in any characterization of backward induction. At any decision node, forward belief in future-high-connectedness

[2] Note that memory connectedness could also be used to interpret issues raised in Piccione and Rubinstein's (1997).

ensures the stability of all game agents at all succeeding decision nodes even if game agents at preceding decision nodes have been deviating from the initial strategy of their reasoning agent. Note that in our framework surprise information precisely consists in deviation from the initial strategy. Epistemic independence is assured by forward belief in future-high-connectedness which, in turn, leads to a behavioral isolation of any surprise information.

In a general sense, note that there is a tension between the sequential stability implicitly assumed to underlie standard accounts of dynamic games and some local breakdown which is needed for epistemic independence. This tension needs to be accounted for in any epistemic characterization of backward induction. In our framework, the notion of connectedness is used to describe this tension: low-connectedness makes explicit the idea that the sequential stability of the initial strategy can break down locally, while forward belief in future-high-connectedness ensures that the effects of such a breakdown indeed remain local. Connectedness thus makes explicit the crucial rigidity with which epistemic independence requires local breakdowns of sequential stability to be treated.

The multiple-self interpretation introduced above yields further insights into the fundamental principle of epistemic independence. In dynamic games, it is natural to depict a player as a multiple-self whose selves are highly connected on all interpretations of underlying connectedness, i.e. selves highly connected in terms of psychological features, memory and sympathy. Upon receiving surprise information, belief revision according to forward belief in future-high-connectedness sets the behavioral connectedness of the deviating opponent game agent to *low*. It is then plausible to claim that this low behavioral connectedness stems from some breakdown in underlying connectedness. However, forward belief in future-high-connectedness also ensures that any succeeding game agents of the respective opponent are assumed to be entirely unaffected by the deviating behavior of the particular preceding supposedly low-connected agent under *any* interpretation of underlying connectedness, such as psychological, sympathetic and memory connectedness. Hence, forward belief in future-high-connectedness reveals that foundations for backward induction have commonly been tacitly assuming a much stronger epistemic independence assumption. Indeed, forward belief in future-high-connectedness is required for any underlying connectedness. Note that assuming such epistemic independence with regards to any underlying connectedness is considerably strong, due to the latter's philosophical foundations. Again, this suggests that the assumption of epistemic independence is much stronger than commonly assumed.

Farther, our framework is capable of clarifying Aumann's (1995) epistemic conditions for backward induction. In his framework, Aumann uses an entirely static epistemic operator that refers to the beginning of the game. Once fixed, the epistemic state of a player concerns a single point in time and does not change. It is hence difficult to account for belief revision in this framework. However, Aumann's key nested epistemic notion of common knowledge of rationality can be interpreted as being equivalent to our concept of common structural belief in rationality. Indeed, rationality refers to a player's initial strategy fixed before the game and rigidity of a belief in a rational initial strategy is thus possible to entertain in our model. However, the belief in the actual choice of the opponents may change at different points in the game. It can hence be claimed that Aumann implicitly endorses some kind of high-connectedness assumption, requiring it to be common knowledge that a player never actually changes his intended initial strategy. This implicit assumption is explicated by our forward belief in future-high-connectedness condition. By understanding strategies as intentions or, more precisely, initial strategies, Aumann is able to obtain backward induction with an entirely static epistemic operator.

5.4 Backward Induction Paradoxes

The so-called backward induction paradoxes have been addressed by, for instance, Selten (1978), Rosenthal (1981) as well as Binmore (1987), and identify games in which backward inductive reasoning leads to rather implausible and counterintuitive strategy choices. In this context, a crucial argument against the plausibility of backward induction criticizes that the reasoning method does not take into account any observed past behavior at all, even when the backward inductive strategy profile is contradicted during actual play. In fact, our framework can be used to juxtapose belief revision patterns in line with such a plausibility requirement, and to contrast them to belief revision policies sufficient for backward induction reasoning according to Theorem 1. Recall that the latter belief revision policies require a player to set the connectedness of a deviating agent to *low* and to maintain belief in the high-connectedness of each of the opponent's future agents. In contrast, belief revision policies in line with the plausibility requirement set the connectedness of all future game agents of the relevant opponent to *low* upon observing an opponent game agent deviate. Thus, the intuition is captured that the respective game agents actually play a strategy different from the initial strategy believed to be chosen by their reasoning agent.

As an illustration of this comparison between these two kinds of belief revision policies for dynamic games with perfect information, consider the dynamic game given in Figure 1. Suppose *Bob* reasons in line with the conditions of Theorem 1 and hence in line with backward induction. In case of him surprisingly observing $Alice_1$ to choose b, he sets her connectedness to *Alice*'s reasoning agent to *low*, while keeping his belief in the high-connectedness of *Alice*'s future game agent $Alice_3$. Alternatively, suppose now that *Bob* when observing $Alice_1$'s deviating move still believes that *Alice*'s reasoning agent has chosen the backward inductive strategy as initial strategy, but that the game agents play a strategy different from the initial strategy. He thus also sets the connectedness between $Alice_3$ and *Alice*'s reasoning agent to *low*. Note that he is free to believe what $Alice_3$ will choose. For instance, if he believes that she will pick e, then he can only optimally select c at his decision node.

We now focus on Rosenthal's (1981) approach to the backward induction paradoxes. On his account, small probabilities of future deviating moves are introduced into dynamic games and interpreted as the players' intersubjective beliefs about future moves of opponents. In our framework, such probabilities can be elucidated by introducing a more fine-grained belief revision on deviating moves. More precisely, we introduce probabilistic beliefs on how likely it is that opponent game agents deviate from their respective initial strategy. These probabilistic beliefs are updated by a player's belief about the underlying connectedness of an opponent, which in turn depends on his beliefs on the opponent's behavioral connectedness. Recall that underlying connectedness describes the degree of continuity of psychological features, memory and sympathy in a multiple-self.

A player could form probabilistic beliefs on future deviating moves of an opponent as follows. Firstly, suppose a player observes a move by an opponent agent that deviated from his respective initial strategy and explains it with the latter's low *behavioral connectedness*. Note that this low behavioral connectedness can be treated as information about the opponent. Secondly, suppose further that a player has a belief about the *underlying connectedness* of the opponent's person. It is then natural to update these beliefs with the behavioral observation. In other words, players can learn about their opponent's character during the game. Thirdly, also suppose that a player entertains beliefs about his opponents' future behavior. Then, it seems reasonable to update the latter beliefs with his beliefs about the respective opponent's underlying connectedness.

Let a specific underlying connectedness be assigned to any *pair* of agents in a player: $c : i \times i \rightarrow \{high, low\}$, where a player i is conceived of as a set of agents according to Definition 3. Such an underlying connectedness function can be interpreted as an opponent's belief about the connectedness between any pair of agents in a player. For instance, within the context of the game given in Figure 1, starting with a natural belief in high-connectedness from sequential stability as well as a belief that *Alice*'s initial strategy is the backward inductive one, upon observing that $Alice_1$ chooses b, backward induction can only be behaviorally maintained for *Bob* by updating his belief about *Alice*'s underlying connectedness function as follows: set the connectedness of any pair of *Alice*'s agents involving $Alice_1$ to *low*, i.e. $c(Alice_s, Alice_1) = low$, $c(Alice_r, Alice_1) = low$, $c(Alice_1, Alice_3) = low$, $c(Alice_1, Alice_s) = low$, $c(Alice_1, Alice_r) = low$, and $c(Alice_3, Alice_1) = low$, while maintaining the belief in the high connectedness between any two other agents, i.e. $c(Alice_s, Alice_r) = high$, $c(Alice_s, Alice_3) = high$, $c(Alice_r, Alice_3) = high$, $c(Alice_r, Alice_s) = high$, $c(Alice_3, Alice_s) = high$, and $c(Alice_3, Alice_r) = high$. Clearly, such a belief revision policy is implausible: believing that $Alice_1$ is low-connected to all other agents while still preserving belief in the high-connectedness between all other pairs of agents seems to completely deny any relevance of $Alice_1$ with regards to *Alice* as a multiple-self.

Consider a more fine-grained underlying connectedness function expressing degrees of connectedness of pairs of agents, namely $c : i \times i \rightarrow [0, 1]$. Such degrees can be natural in the context of interpretations of connectedness from theories of personal identity over time. For example, according to psychological connectedness, each agent could be interpreted as a set of preferences. The degree of connectedness then measures the rate of preference change. Similarly, memory connectedness can be seen as a matter of degrees rather than binary. Interpreting the more-fine grained connectedness function c with theories of personal identity over time further illustrates that backward induction can be argued to be very demanding in terms of maintaining beliefs in high-connectedness of future game agents of opponents. Using such a more fine-grained underlying connectedness function, the following belief revision upon receiving surprise information seems plausible. The supposedly low behavioral connectedness of the deviating game agent $Alice_1$ induces Bob to believe that there is some agent of $Alice$ to which $Alice_1$'s connectedness is strictly less than 1. In other words, some failure of the underlying connectedness has to be assumed in order to explain the low behavioral connectedness. Then, such revised beliefs in an underlying connectedness function can be used to update beliefs about future moves of an opponent. More specifically, underlying connectedness can determine a player's probabilistic beliefs about how likely an opponent game agent will deviate from the initial strategy at future nodes. Conditionalizing on the fact there is some agent of $Alice$ to which $Alice_1$'s connectedness is strictly less than 1, plausible updating rules render Bob's probabilistic beliefs in future deviation of any of $Alice$'s game agents strictly positive, since each agent has some relevance to his respective player as a multiple-self. Intuitively, upon believing that there is at least some agent-pair in an opponent which is not perfectly connected, the respective player's belief about the future deviation of his opponent's game agents will be strictly positive, as future game agents may also be disposed to deviate from the reasoning agent's initial strategy as exhibited by the particular game agent that has already deviated.

Further plausible constraints on such updating patterns can be introduced. For instance, it is possible that under the interpretation of psychological connectedness, the underlying connectedness changes more drastically than under the interpretation of memory connectedness. Consider a preference change of one game agent whose psychological connectedness is thus low. If all other agents' preferences remain stable, then there will be a low-connectedness between the respective game agent and all other agents. However, in the context of memory connectedness, it could be the case that one game agent has forgotten the initial strategy, while all other agents still remember the initial strategy well. Therefore, it is at least plausible to require monotonicity to be respected in updating beliefs about future deviation for psychological connectedness. With different interpretations, the breakdown of connectedness can be more or less wide in scope in terms of how many agents are affected. Which or whether all of these interpretations are endorsed depends on how realistic a model of the decision-maker is intended.

In our framework, it can thus be argued that backward induction reasoning is implausible, when underlying connectedness is interpreted as a belief about the opponent's character and probabilistic beliefs about future deviation of opponents are updated on the basis of beliefs about underlying connectedness.

5.5 Trembling Hand

It is possible to interpret Selten's (1975) idea of a perturbed game with connectedness as understood in our framework in a natural way. The claim that a deviating move is due to the respective player exhibiting a trembling hand, and thus making a slight mistake by picking an irrational action with small probability, can be expressed and explained in our model. The agents of a player are assumed to be highly connected and play in line with the initial strategy with almost certainty, yet with small probability their underlying connectedness is low and subsequent behavior deviates. In other words, a given game agent might - despite it being assumed to be very unlikely - tremble in implementing the intended initial strategy of his player's reasoning agent. Such trembles can be used as explanations for observed deviations from an opponent's supposed initial strategy in belief revision policies. More precisely, whenever a player who believes in the high-connectedness of the game agents of each of his opponents is surprised by a move of a game agent which contradicts the initial strategy he believes the respective opponent's reasoning agent to have chosen, he can then separate that game agent. Indeed he can only assign low-connectedness to that particular game agent, while keeping fixed the high-connectedness of the respective opponent's other game agents. Such isolated behavior of this given deviating game agent is explained as a mistake on the agent's part.

Note that such a specific trembling hand vindication of deviating behavior corresponds to a particular belief concerning the underlying connectedness of the relevant game agent. Intuitively, the trembling hand is a physical metaphor for the failure to complete a given task, despite having had appropriate dispositions to perform it. In terms of underlying connectedness, it is possible to interpret a trembling hand with low sympathetic connectedness, while at the same time psychological as well as memory connectedness are high. Hence, the deviating game agent is now supposed to be lowly connected to the player as a multiple-self: even though he has the same preferences and perfect memory, he somehow slips and makes a mistake. Note that such belief revision policies are close to the ones sufficient for backward induction and far from the supposedly more plausible ones with regards to the backward induction paradoxes in terms of their general intuition. Lexicographically speaking, whenever a surprise move of some game agent contradicting the supposed initial strategy of his respective reasoning agent is observed, the state in which the deviating agent is lowly connected and others are highly connected is deemed infinitely more likely than the state where the player's future agents are also lowly connected to the respective player. The key to the construction and comparisons of such belief patterns is our notion of initial strategy, which can be contrasted with the same player's actual strategy, and hence belief about initial choice can be juxtaposed with belief about actual choice.

As an illustration of the idea of a trembling hand in the context of our framework, consider the dynamic game given in Figure 1. Suppose *Bob* initially believes all game agents of *Alice* to be high-connected and at Bob_2 that *Alice*'s initial strategy is backward inductive one af. However, at Bob_2 he then has

has to accommodate the surprise information that $Alice_1$ has actually chosen b. Explaining this deviating behavior with an exceptional mistake incurred by $Alice_1$ in implementing $Alice_r$'s plan, Bob sets the connectedness of game agent $Alice_1$ to low, yet preserves his belief in the high-connectedness of $Alice$'s future game agent $Alice_3$. His unique optimal choice is hence given by d.

6 Conclusion

By rendering transparent relevant yet usually neglected processes linked to dynamic games we clarify their inherent dynamics. We analyze the sequential structure of dynamic games with a three-stage account, which defines player and strategy relative to these dynamic stages. A sequential stability assumption underlying the standard extensive form model of dynamic games is made explicit in our account. To describe reasoning in dynamic games, a more general epistemic model is proposed that is capable of formalizing the notion of agent connectedness. Such an enriched framework sheds light on backward induction reasoning. Formally, we provide sufficient conditions for backward induction in terms of connectedness. Conceptually, the essence of backward induction can be explicated, since surprise information is explained with low-connectedness of the deviating agent. Also, the epistemic independence assumption underlying any foundation of backward induction can be shown to be considerably stronger than usually assumed. Our framework makes explicit that any underlying connectedness of players as multiple-selves has tacitly been assumed to be high.

In a general sense, our frameworks provides adequate foundations for interpreting the sequential structure of dynamic games in temporal terms. In particular, defining a player as a set of agents enables a more realistic interpretation of decision-makers in dynamic games. Using the multiple-self model of personal identity over time also provides richer descriptions of players, for instance, with regards to psychological, sympathetic and memory connectedness. Hence, our framework is especially relevant for the social sciences, where players should be interpreted as persons existing over time.

Finally, the framework proposed here could also be employed to shed light on the sequential structure and dynamics of games of imperfect information as well as to to clarify corresponding reasoning and solution concepts. It would be of particular interest to search for sequential stability requirements for forward induction reasoning in terms of agent connectedness. Intuitively, actual choice of a game agent should then be believed to be highly relevant to actual choice of the respective player's future game agents.

Acknowledgment

We would like to thank Johan van Benthem, Richard Bradley, Adam Brandenburger, Jacques Duparc, Eric Pacuit and Andrés Perea as well as audiences at the London School of Economics (LSE) and the Institute for Logic, Language and Computation (ILLC) at the University of Amsterdam.

References

Ainslie, G., 1992. *Picoeconomics.* Cambridge University Press.

Aumann, R. J., 1976. Agreeing to Disagree. *The Annals of Statistics*, 4, 1236–1239.

Aumann, R. J., 1995. Backward Induction and Common Knowledge of Rationality. *Games and Economic Behavior*, 8, 6–19.

Battigalli, P., Siniscalchi, M., 2002. Strong Belief and Forward Induction Reasoning. *Journal of Economic Theory*, 106, 356–391.

Binmore, K., 1987. Modeling rational players I, *Economics and Philosophy*, 3, 179–214.

Brandenburger, A., Friedenberg, A., Keisler, H. J., 2008, Admissibility in Games. *Econometrica*, 76, 307352.

Elster, J., 1986. *The Multiple Self.* Cambridge University Press.

Halpern, J., 2001. Substantive Rationality and Backward Induction. *Games and Economic Behavior*, 37, 425–435.

Harsanyi, J. C., 1967-68. Games of Incomplete Information played by "Bayesian Players". Part I, II, III. *Management Science*, 14, 159–182, 320–334, 486–502.

Hume, D., 1739. *A Treatise of Human Nature.* Clarendon, Oxford.

Kuhn, H. W., 1953. Extensive games and the problem of information. *Annals of Mathematics Studies*, 28, 193–216.

Locke, J., 1694. *An Essay concerning Human Understanding.* Book II, chapter XXVII. Clarendon, Oxford.

Noonan, H. W., 2003. *Personal Identity.* Routledge, London.

Parfit, D., 1984. *Reasons and Persons.* Clarendon, Oxford.

Perea, A., 2007. Epistemic Conditions for Backward Induction: An Overview. In: *Interactive Logic Proceedings of the 7th Augustus de Morgan Workshop, London. Texts in Logic and Games 1.* Amsterdam University Press, pp. 159–193.

Perea, A., 2008. Minimal Belief Revision leads to Backward Induction. *Mathematical Social Sciences*, 56, 1–26.

Piccione, M., Rubinstein, A., 1997. On the Interpretation of Decision Problems with Imperfect Recall. *Games and Economic Behavioir*, 20, 3–24.

Raymond, M., Barresi, J. (Eds.), 2003. *Personal Identity.* Blackwell, Oxford.

Reny, P. W., 1992. Rationality in Extensive-Form Games. *Journal of Economic Perspectives*, 6, 103–118.

Reny, P. W., 1993. Common Belief and the Theory of Games with Perfect Information. *Journal of Economic Theory*, 59, 257–274.

Rosenthal, R.W., 1981. Games of Perfect Information, Predatory Pricing and the Chain-Store Paradox. *Journal of Economic Theory*, 25, 92–100.

Schechtmann, M., 2001. Empathic Access. *Philosophical Explorations*, 4, 95–111.

Selten, R., 1975. Reexamination of the perfectness concept of equilibrium in extensive games. *International Journal of Game Theory*, 4, 25–55.

Selten, R., 1978. The Chain Store Paradox. *Theory and Decision*, 9, 121–159.

Shoemaker, S., Swinburne, R.G., 1984. *Personal Identity.* Blackwell, Oxford.

Siniscalchi, M., 2008. Epistemic Game Theory: Beliefs and Types. In: *The New Palgrave Dictionary of Economics*, Steven N. Durlauf and Lawrence E. Blume (Eds.), Palgrave Macmillan.

Stalnaker, R., 1998. Belief Revision in Games: Forward and Backward Induction. *Mathematical Social Sciences*, 36, 31–56.

Tan, T. C., Werlang, S. R. C., 1988. The Bayesian foundation of Solution Concepts of Games. *Journal of Economic Theory*, 45, 370–391.