Chapter 2 in: J. van Benthem, S. Ghosh and R. Verbrugge (Eds), Models of Strategic Reasoning, LNCS 8972, Springer, 2015, pp. 34–62.

Reasoning about strategies and rational play in dynamic games

Giacomo Bonanno *

Department of Economics, University of California, Davis, USA
gfbonanno@ucdavis.edu

Abstract

We discuss the issues that arise in modeling the notion of common belief of rationality in epistemic models of dynamic games, in particular at the level of interpretation of strategies. A strategy in a dynamic game is defined as a function that associates with every information set a choice at that information set. Implicit in this definition is a set of counterfactual statements concerning what a player would do at information sets that are not reached, or a belief revision policy concerning behavior at information sets that are ruled out by the initial beliefs. We discuss the role of both objective and subjective counterfactuals in attempting to flesh out the interpretation of strategies in epistemic models of dynamic games.

1 Introduction

Game theory provides a formal language for the representation of interactive situations, that is, situations where several "entities" - called players - take actions that affect each other. The nature of the players varies depending on the

*I am grateful to Sonja Smets for presenting this chapter at the Workshop on Modeling Strategic Reasoning (Lorentz Center, Leiden, February 2012) and for offering several constructive comments. I am also grateful to two anonymous reviewers and to the participants in the workshop for many useful comments and suggestions.

context in which the game theoretic language is invoked: in evolutionary biology (see, for example, Smith (1982)) players are non-thinking living organisms;[1] in computer science (see, for example, Shoham and Leyton-Brown (2008)) players are artificial agents; in behavioral game theory (see, for example, Camerer (2003)) players are "ordinary" human beings, etc. Traditionally, however, game theory has focused on interaction among intelligent, sophisticated and rational individuals. For example, Aumann describes game theory as follows:

> "Briefly put, game and economic theory are concerned with the interactive behavior of *Homo rationalis* - rational man. *Homo rationalis* is the species that always acts both purposefully and logically, has well-defined goals, is motivated solely by the desire to approach these goals as closely as possible, and has the calculating ability required to do so." (Aumann (1985), p. 35)

This chapter is concerned with the traditional interpretation of game theory, in particular, with what is known as the *epistemic foundation program*, whose aim is to characterize, for any game, the behavior of rational and intelligent players who know the structure of the game and the preferences of their opponents and who recognize each other's rationality and reasoning abilities. The fundamental problem in this literature is to answer the following two questions: (1) under what circumstances can a player be said to be rational? and (2) what does 'mutual recognition' of rationality mean? While there seems to be agreement in the literature that 'mutual recognition' of rationality is to be interpreted as 'common belief' of rationality, the issue of what it means to say that a player is rational is not settled. Everybody agrees that the notion of rationality involves two ingredients: choice and beliefs. However, the precise nature of their relationship involves subtle issues which will be discussed below, with a focus on dynamic games. We shall restrict attention to situations of complete information, which are defined as situations where the game being played is common knowledge among the players.[2]

There is a bewildering collection of claims in the literature concerning the implications of rationality in dynamic games with perfect information: Aumann (1995) proves that common *knowledge* of rationality implies the backward

[1] Evolutionary game theory has been applied not only to the analysis of animal and insect behavior but also to studying the "most successful strategies" for tumor and cancer cells (see, for example, Gerstung et al. (2011)).

[2] On the other hand, in a situation of *incomplete* information at least one player lacks knowledge of some of the aspects of the game, such as the preferences of her opponents, or the actions available to them, or the possible outcomes, etc.

induction solution, Ben-Porath (1997) and Stalnaker (1998) prove that common *belief / certainty* of rationality is *not* sufficient for backward induction, Samet (1996) proves that what is needed for backward induction is common *hypothesis* of rationality, Feinberg (2005) shows that common *confidence* of rationality logically contradicts the knowledge implied by the structure of the game, etc. The purpose of this chapter is not to review this literature[3] but to highlight some of the conceptual issues that have emerged.

In Section 2 we start with a brief exposition of one of the essential components of a definition of rationality, namely the concept of belief, and we review the notions of a model of a game and of rationality in the context of simultaneous games. We also discuss the role of counterfactuals in the analysis of simultaneous games. In the context of dynamic games there is a new issue that needs to be addressed, namely what it means to choose a strategy and what the proper interpretation of strategies is. This is addressed in Section 3 where we also discuss the subtle issues that arise when attempting to define rationality in dynamic games.[4] In Section 4 we turn to the topic of belief revision in dynamic games and explore the use of subjective counterfactuals in the analysis of dynamic games with perfect information. Section 5 concludes.

The formalism is introduced gradually throughout the chapter and only to the extent that is necessary to give precise content to the concepts discussed. For the reader's convenience a table in the Appendix summarizes the notations used and the corresponding interpretations.

The analysis is carried out entirely from a semantic perspective.[5]

2 Belief, common belief and models of games

For simplicity, we shall restrict attention to a qualitative notion of belief, thus avoiding the additional layer of complexity associated with probabilistic or graded beliefs.

Definition 2.1. An *interactive belief structure* (or *multi-agent Kripke structure*) is a tuple $\langle N, \Omega, \{\mathcal{B}_i\}_{i\in N}\rangle$ where N is a finite set of *players*, Ω is a set of *states* and, for every player $i \in N$, $\mathcal{B}_i$ is a binary relation on Ω representing *doxastic accessibility*: the interpretation of $\omega\mathcal{B}_i\omega'$ is that at state ω player i considers state ω' possible.

[3] Surveys of the literature on the epistemic foundations of game theory can be found in Battigalli and Bonanno (1999), Brandenburger (2007), Dekel and Gul (1997), Perea (2007; 2012).

[4] The notion of rationality in dynamic games is also discussed in Perea (2014).

[5] For a syntactic analysis see Bonanno (2008; forthcomingb), Clausing (2003; 2004), de Bruin (2010), van Benthem (2011). See also Pacuit (2014).

We denote by $\mathcal{B}_i(\omega)$ the set of states that are compatible with player i's beliefs at state ω,[6] that is,

$$\mathcal{B}_i(\omega) = \{\omega' \in \Omega : \omega \mathcal{B}_i \omega'\}. \tag{1}$$

We assume that each $\mathcal{B}_i$ is serial ($\mathcal{B}_i(\omega) \neq \emptyset$, $\forall \omega \in \Omega$), transitive (if $\omega' \in \mathcal{B}_i(\omega)$ then $\mathcal{B}_i(\omega') \subseteq \mathcal{B}_i(\omega)$) and euclidean (if $\omega' \in \mathcal{B}_i(\omega)$ then $\mathcal{B}_i(\omega) \subseteq \mathcal{B}_i(\omega')$). Seriality captures the notion of consistency of beliefs, while the last two properties correspond to the notions of positive and negative introspection of beliefs.[7]

Subsets of Ω are called *events*. We shall use E and F as variables for events. Associated with the binary relation $\mathcal{B}_i$ is a *belief operator* on events $\mathbb{B}_i : 2^\Omega \to 2^\Omega$ defined by

$$\mathbb{B}_i E = \{\omega \in \Omega : \mathcal{B}_i(\omega) \subseteq E\}. \tag{2}$$

Thus $\mathbb{B}_i E$ is the event that player i believes E.[8]

Figure 1 shows an interactive belief structure with two players, where each relation $\mathcal{B}_i$ is represented by arrows: $\omega' \in \mathcal{B}_i(\omega)$ if and only if there is an arrow, for player i, from ω to ω'. Thus, in Figure 1, we have that $\mathcal{B}_1 = \{(\alpha,\alpha),(\beta,\gamma),(\gamma,\gamma)\}$ and $\mathcal{B}_2 = \{(\alpha,\alpha),(\beta,\alpha),(\gamma,\gamma)\}$, so that, for example, $\mathcal{B}_1(\beta) = \{\gamma\}$ while $\mathcal{B}_2(\beta) = \{\alpha\}$. In terms of belief operators, in this structure we have that, for instance, $\mathbb{B}_1\{\gamma\} = \{\beta,\gamma\}$, that is, at both states β and γ Player 1 believes event $\{\gamma\}$, while $\mathbb{B}_2\{\gamma\} = \{\gamma\}$, so that Player 2 believes event $\{\gamma\}$ only at state γ.

Let $\mathcal{B}^*$ be the transitive closure of $\bigcup_{i\in N} \mathcal{B}_i$ [9] and define the corresponding operator $\mathbb{B}^* : 2^\Omega \to 2^\Omega$ by

$$\mathbb{B}^* E = \{\omega \in \Omega : \mathcal{B}^*(\omega) \subseteq E\}. \tag{3}$$

$\mathbb{B}^*$ is called the *common belief* operator and when $\omega \in \mathbb{B}^* E$ then at state ω every player believes E and every player believes that every player believes E, and so on, *ad infinitum*.

[6] Thus $\mathcal{B}_i$ can also be viewed as a function from Ω into 2^Ω (the power set of Ω). Such functions are called *possibility correspondences* (or information functions) in the game-theoretic literature.

[7] For more details see the survey in Battigalli and Bonanno (1999).

[8] In modal logic belief operators are defined as syntactic operators on formulas. Given a (multi-agent) Kripke structure, a model based on it is obtained by associating with every state an assignment of truth value to every atomic formula (equivalently, by associating with every atomic formula the set of states where the formula is true). Given an arbitrary formula ϕ, one then stipulates that, at a state ω, the formula $B_i\phi$ (interpreted as 'agent i believes that ϕ') is true if and only if ϕ is true at every state $\omega' \in \mathcal{B}_i(\omega)$ (that is, $\mathcal{B}_i(\omega)$ is a subset of the truth set of ϕ). If event E is the truth set of formula ϕ then the event $\mathbb{B}_i E$ is the truth set of the formula $B_i\phi$.

[9] That is, $\omega' \in \mathcal{B}^*(\omega)$ if and only if there is a sequence $\langle \omega_1, ..., \omega_m \rangle$ in Ω and a sequence $\langle j_1, ..., j_{m-1} \rangle$ in N such that (1) $\omega_1 = \omega$, (2) $\omega_m = \omega'$ and (3) for all $k = 1, ..., m-1$, $\omega_{k+1} \in \mathcal{B}_{j_k}(\omega_k)$.

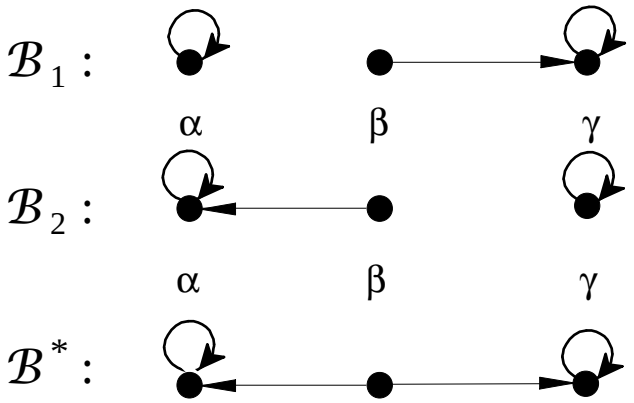

Figure 1: An interactive belief structure

Figure 1 shows the relation $\mathcal{B}^*$ (the transitive closure of $\mathcal{B}_1 \cup \mathcal{B}_2$): in this case we have that, for example, $\mathbb{B}^*\{\gamma\} = \{\gamma\}$ and thus $\mathbb{B}_1\mathbb{B}^*\{\gamma\} = \{\beta, \gamma\}$, that is, event $\{\gamma\}$ is commonly believed only at state γ, but at state β Player 1 erroneously believes that it is common belief that $\{\gamma\}$ is the case.[10]

When the relations $\mathcal{B}_i$ ($i \in N$) are also assumed to be reflexive ($\omega \in \mathcal{B}_i(\omega)$, $\forall \omega \in \Omega$), then they become equivalence relations and thus each $\mathcal{B}_i$ gives rise to a partition of Ω. In partitional models, beliefs are necessarily correct and one can speak of *knowledge* rather than belief. As Stalnaker (1996) points out, it is methodologically preferable to carry out the analysis in terms of (possibly erroneous) beliefs and then - if desired - add further conditions that are sufficient to turn beliefs into knowledge. The reason why one should not start with the assumption of necessarily correct beliefs (that is, reflexivity of the $\mathcal{B}_i$'s) is that this assumption has strong intersubjective implications:

"The assumption that Alice believes (with probability one) that Bert

[10] As can be seen from Figure 1, the common belief relation $\mathcal{B}^*$ is not necessarily euclidean, despite the fact that the $\mathcal{B}_i$'s are euclidean. In other words, in general, the notion of common belief does not satisfy negative introspection (although it does satisfy positive introspection). It is shown in Bonanno and Nehring (2000) that negative introspection of common belief holds if and only if no agent has erroneous beliefs about what is commonly believed.

> believes (with probability one) that the cat ate the canary tells us nothing about what Alice believes about the cat and the canary themselves. But if we assume instead that Alice knows that Bert knows that the cat ate the canary, it follows, not only that the cat in fact ate the canary, but that Alice knows it, and therefore believes it as well." (Stalnaker (1996), p. 153.)

One can express locally (that is, at a state ω) the properties of knowledge by means of the double hypothesis that, at that state, at least one player has correct beliefs (for some $i \in N$, $\omega \in \mathcal{B}_i(\omega)$) and that there is common belief that nobody has erroneous beliefs (for all $\omega' \in \mathcal{B}^*(\omega)$ and for all $i \in N$, $\omega' \in \mathcal{B}_i(\omega')$).[11] Adding such hypotheses introduces strong forms of agreements among the players (see Bonanno and Nehring (1998)) and is, in general, not realistic.

Interactive belief structures can be used to model particular contexts in which a game is played. Let us take, as a starting point, strategic-form games (also called normal-form games), where players make their choices simultaneously (an example is a sealed-bid auction).[12]

Definition 2.2. A *strategic-form game with ordinal payoffs* is a tuple $\langle N, \{S_i, \succsim_i\}_{i \in N} \rangle$ where N is a set of *players* and, for every $i \in N$, S_i is a set of choices or *strategies* available to player i and $\succsim_i$ is i's preference relation over the set of *strategy profiles* $S = \underset{i \in N}{\times} S_i$.[13]

[11] This is a local version of knowledge (defined as true belief) which is compatible with the existence of other states where some or all players have erroneous beliefs (see Bonanno and Nehring (1998), in particular Definition 2 on page 9 and the example of Figure 2 on page 6). Note that philosophical objections have been raised to defining knowledge as true belief; for a discussion of this issue see, for example, Stalnaker (2006).

[12] Strategic-form games can also be used to represent situations where players move sequentially, rather than simultaneously. This is because, as discussed later, strategies in such games are defined as complete, contingent plans of action. However, the choice of a strategy in a dynamic game is thought of as being made before the game begins and thus the strategic-form representation of a dynamic game can be viewed as a simultaneous game where all the players choose their strategies simultaneously before the game is played.

[13] A preference relation over a set S is a binary relation $\succsim$ on S which is complete or connected (for all $s, s' \in S$, either $s \succsim s'$ or $s' \succsim s$, or both) and transitive (for all $s, s', s'' \in S$, if $s \succsim s'$ and $s' \succsim s''$ then $s \succsim s''$). We write $s \succ s'$ as a short-hand for $s \succsim s'$ and $s' \not\succsim s$ and we write $s \sim s'$ as a short-hand for $s \succsim s'$ and $s' \succsim s$. The interpretation of $s \succsim_i s'$ is that player i considers s to be at least as good as s', while $s \succ_i s'$ means that player i prefers s to s' and $s \sim_i s'$ means that she is indifferent between s and s'. The interpretation is that there is a set Z of possible outcomes over which every player has a preference relation. An outcome function $o : S \to Z$ associates an outcome with every strategy profile, so that the preference relation over Z induces a preference relation over S.

We shall throughout focus on ordinal preferences (rather than cardinal preferences with associated expected utility comparisons)[14] for two reasons: (1) since the game is usually hypothesized to be common knowledge among the players, it seems far more realistic to assume that each player knows the ordinal rankings of her opponents rather than their full attitude to risk (represented by a cardinal utility function) and (2) our aim is to point out some general conceptual issues, which are independent of the notion of expected utility.

The definition of a strategic-form game specifies the choices available to the players and what motivates those choices (their preferences over the possible outcomes); however, it leaves out an important factor in the determination of players' choices, namely what they believe about the other players. Adding a specification of the players' beliefs determines the context in which a particular game is played and this can be done with the help of an interactive belief structure.

Definition 2.3. Fix a strategic-form game $G = \langle N, \{S_i, \succsim_i\}_{i \in N} \rangle$. A *model of* G is a tuple $\langle N, \Omega, \{\mathcal{B}_i\}_{i \in N}, \{\sigma_i\}_{i \in N} \rangle$, where $\langle N, \Omega, \{\mathcal{B}_i\}_{i \in N} \rangle$ is an interactive belief structure (see Definition 2.1) and, for every $i \in N$, $\sigma_i : \Omega \to S_i$ is a function that assigns to each state ω a strategy $\sigma_i(\omega) \in S_i$ of player i.

Let $\sigma(\omega) = (\sigma_i(\omega))_{i \in N}$ denote the strategy profile associated with state ω. The function $\sigma : \Omega \to S$ gives content to the players' beliefs. If $\omega \in \Omega$, $x \in S_i$ and $\sigma_i(\omega) = x$ then the interpretation is that at state ω player i "chooses" strategy x. The exact meaning of 'choosing' is not elaborated further in the literature: does it mean that player i *has actually played* x, or that she is *committed to playing* x, or that x is the *output of her deliberation process*? Whatever the answer, the assumption commonly made in the literature is that player i has correct beliefs about her chosen strategy, that is, she chooses strategy x if and only if she believes that her chosen strategy is x. This can be expressed formally as follows. For every $x \in S_i$, let $[\sigma_i = x]$ be the event that player i chooses strategy x, that is, $[\sigma_i = x] = \{\omega \in \Omega : \sigma_i(\omega) = x\}$. Then the assumption is that

$$[\sigma_i = x] = \mathbb{B}_i [\sigma_i = x]. \tag{4}$$

We will return to this assumption later on, in our discussion of dynamic games. Figure 2 shows a strategic-form game in the form of a table, where the preference relation $\succsim_i$ of player i is represented numerically by an ordinal *utility function* $u_i : S \to \mathbb{R}$, that is, a function satisfying the property that $u_i(s) \geq u_i(s')$

[14] Cardinal utility functions are also called Bernoulli utility functions or von Neumann-Morgenstern utility functions.

		Player 2	
		l	r
Player 1	t	2 , 1	0 , 0
	b	1 , 2	1 , 2

Figure 2: A strategic form game

if and only if $s \gtrsim_i s'$. In each cell of the table the first number is the utility of Player 1 and the second number the utility of Player 2. A model of this game can be obtained by adding to the interactive belief frame of Figure 1 the following strategy assignments:

$$\begin{array}{l} \sigma_1(\alpha) = b, \;\; \sigma_1(\beta) = \sigma_1(\gamma) = t \\ \sigma_2(\alpha) = \sigma_2(\beta) = r, \;\; \sigma_2(\gamma) = l. \end{array} \tag{5}$$

How can rationality be captured in a model? Consider the following - rather weak - definition of rationality: player i is rational at state $\hat{\omega}$ if - given that she chooses the strategy $\hat{s}_i \in S_i$ at state $\hat{\omega}$ (that is, given that $\sigma_i(\hat{\omega}) = \hat{s}_i$) - there is no other strategy $s_i \in S_i$ which player i believes, at state $\hat{\omega}$, to be better (that is, to yield a higher payoff) than $\hat{s}_i$. This can be stated formally as follows. First of all, for every state ω, denote by $\sigma_{-i}(\omega)$ the strategy profile of the players other than i, that is, $\sigma_{-i}(\omega) = (\sigma_1(\omega), ..., \sigma_{i-1}(\omega), \sigma_{i+1}(\omega), ..., \sigma_n(\omega))$ (where n is the number of players). Then (recall that - since $\sigma_i(\hat{\omega}) = \hat{s}_i$ - by (4) $\sigma_i(\omega) = \hat{s}_i$, for all $\omega \in \mathcal{B}_i(\hat{\omega})$):

$$\begin{array}{l} \text{Player } i \text{ is } \textit{rational at } \hat{\omega} \text{ if, } \forall s_i \in S_i, \text{ it is not the case} \\ \text{that, } \forall \omega \in \mathcal{B}_i(\hat{\omega}), \;\; u_i\left(s_i, \sigma_{-i}(\omega)\right) > u_i\left(\hat{s}_i, \sigma_{-i}(\omega)\right) \\ (\text{where } \hat{s}_i = \sigma_i(\hat{\omega})). \end{array} \tag{6}$$

Equivalently, let $[u_i(s_i) > u_i(\hat{s}_i)] = \{\omega \in \Omega : u_i\left(s_i, \sigma_{-i}(\omega)\right) > u_i\left(\hat{s}_i, \sigma_{-i}(\omega)\right)\}$. Then

$$\begin{array}{l} \text{Player } i \text{ is } \textit{rational at } \hat{\omega} \text{ if, } \forall s_i \in S_i, \; \hat{\omega} \notin \mathbb{B}_i\left[u_i(s_i) > u_i(\hat{s}_i)\right] \\ (\text{where } \hat{s}_i = \sigma_i(\hat{\omega})). \end{array} \tag{7}$$

For example, in the model of the strategic-form game of Figure 2 obtained by adding to the interactive belief structure of Figure 1 the strategy assignments given above in (5), we have that both players are rational at every state and thus there is common belief of rationality at every state. In particular, there is common belief of rationality at state β, even though the strategy profile actually chosen there is (t, r) (with payoffs $(0, 0)$) and each player would do strictly better with a different choice of strategy. Note also that, in this model, at every state it is common belief between the players that each player has correct beliefs,[15] although at state β neither player does in fact have correct beliefs.

It is well known that, in any model of any finite strategic-form game, a strategy profile $s = (s_i)_{i \in N}$ is compatible with common belief of rationality if and only if, for every player i, the strategy s_i survives the iterated deletion of strictly dominated strategies.[16]

What is the conceptual content of the definition given in (7)? It is widely claimed that the notion of rationality involves the use of counterfactual reasoning. For example, Aumann writes:

> "[O]ne really cannot discuss rationality, or indeed decision making, without substantive conditionals and counterfactuals. Making a decision means choosing among alternatives. Thus one must consider hypothetical situations - what would happen if one did something different from what one actually does. [. . .] In interactive decision making - games - you must consider what other people would do if you did something different from what you actually do." (Aumann (1995), p. 15)

Yet the structures used so far do not incorporate the tools needed for counterfactual reasoning. The definition of rationality given in (7) involves comparing the payoff of a strategy different from the one actually chosen with the payoff of the chosen strategy. Can this counterfactual be made explicit?

First we review the standard semantics for counterfactuals.[17]

[15] That is, $\forall \omega \in \Omega$, $\forall \omega' \in \mathcal{B}^*(\omega)$, $\omega' \in \mathcal{B}_1(\omega')$ and $\omega' \in \mathcal{B}_2(\omega')$.

[16] Thus, if at a state ω there is common belief of rationality then, for every player i, $\sigma_i(\omega)$ survives the iterated deletion of strictly dominated strategies. For more details on this result, which originates in Bernheim (1984) and Pearce (1984), and relevant references, see Battigalli and Bonanno (1999), Bonanno (forthcomingb), Dekel and Gul (1997), Perea (2012).

[17] For an extensive discussion see Halpern (1999b). In the game-theoretic literature (see, for example Board (2006) and Zambrano (2004)) a simpler approach is often used (originally introduced by Stalnaker (1968)) where $f(\omega, E)$ is always a singleton.

Definition 2.4. Given a set of states Ω and a set $\mathcal{E} \subseteq 2^{\Omega} \setminus \varnothing$ of events, interpreted as admissible hypotheses, a *counterfactual selection function* is a function $f : \Omega \times \mathcal{E} \to 2^{\Omega}$ that satisfies the following properties: $\forall \omega \in \Omega$, $\forall E, F \in \mathcal{E}$,

$$\begin{array}{ll} 1. & f(\omega, E) \neq \varnothing. \\ 2. & f(\omega, E) \subseteq E. \\ 3. & \text{If } \omega \in E \text{ then } f(\omega, E) = \{\omega\}. \\ 4. & \text{If } E \subseteq F \text{ and } f(\omega, F) \cap E \neq \varnothing \text{ then } f(\omega, E) = f(\omega, F) \cap E. \end{array} \tag{8}$$

The event $f(\omega, E)$ is interpreted as "the set of states closest to ω where E is true". Condition 1 says that there indeed exist states closest to ω where E is true (recall that if $E \in \mathcal{E}$ then $E \neq \varnothing$). Condition 2 is a consistency condition that says that the states closest to ω where E is true are indeed states where E is true. Condition 3 says that if E is true at ω then there is only one state closest to ω where E is true, namely ω itself. Condition 4 says that if E implies F and some closest F-states to ω are in E, then the closest E-states to ω are precisely those states in E that are also the closest F-states to ω.[18]

Given a hypothesis $E \in \mathcal{E}$ and an event $F \subseteq \Omega$, a counterfactual statement of the form "if E were the case then F would be the case", which we denote by $E \rightrightarrows F$, is considered to be true at state ω if and only if $f(\omega, E) \subseteq F$, that is, if F is true in the closest states to ω where E is true. Thus, one can define the operator $\rightrightarrows : \mathcal{E} \times 2^{\Omega} \to 2^{\Omega}$ as follows:

$$E \rightrightarrows F = \{\omega \in \Omega : f(\omega, E) \subseteq F\}. \tag{9}$$

Adding a counterfactual selection function to an interactive belief structure allows one to consider complex statements of the form "if E were the case then player i would believe F" (corresponding to the event $E \rightrightarrows \mathbb{B}_i F$), or "player i believes that if E were the case then F would be the case" (corresponding to $\mathbb{B}_i(E \rightrightarrows F)$), or "Player 1 believes that if E were the case then Player 2 would believe F" (corresponding to $\mathbb{B}_1(E \rightrightarrows \mathbb{B}_2 F)$), etc.

Now, returning to models of strategic-form games and the definition of rationality given in (7), the addition of a counterfactual selection function to a model allows one to compare player i's payoff at a state $\hat{\omega}$, where she has chosen strategy $\hat{s}_i$, with her payoff at the states closest to $\hat{\omega}$ where she chooses a strategy

[18] When $\mathcal{E}$ coincides with $2^{\Omega} \setminus \varnothing$, Condition 4 implies that, for every $\omega \in \Omega$, there exists a complete and transitive "closeness to ω" binary relation $\preceq_{\omega}$ on Ω such that $f(\omega, E) = \{\omega' \in E : \omega' \preceq_{\omega} x, \forall x \in E\}$ (see Theorem 2.2 in Suzumura (1983)) thus justifying the interpretation suggested above: $\omega_1 \preceq_{\omega} \omega_2$ is interpreted as 'state ω_1 is closer to ω than state ω_2 is' and $f(\omega, E)$ is the set of states in E that are closest to ω.

$s_i \neq \hat{s}_i$. Implicit in (7) is the assumption that in those counterfactual states player i's beliefs about her opponents' choices are the same as in $\hat{\omega}$. This is an assumption: it may be a sensible one to make (indeed Stalnaker Stalnaker (1998; 1999) argues that it would be conceptually wrong *not* to make this assumption) but nonetheless it may be worthwhile bringing it to light in a more complete analysis where counterfactuals are explicitly modeled. Within the context of strategic-form games, this is done in Board (2006) and Zambrano (2004), where counterfactuals are invoked explicitly in the definition of rationality.[19]

3 Models of dynamic games

In dynamic games (also called extensive-form games) players make choices sequentially, having some information about the moves previously made by their opponents. If information is partial, the game is said to have *imperfect information*, while the case of full information is referred to as *perfect information*. We shall focus on perfect-information games, which are defined as follows. If A is a set, we denote by A^* the set of finite sequences in A. If $h = \langle a_1, ..., a_k \rangle \in A^*$ and $1 \leq j \leq k$, the sequence $\langle a_1, ..., a_j \rangle$ is called a *prefix* of h. If $h = \langle a_1, ..., a_k \rangle \in A^*$ and $a \in A$, we denote the sequence $\langle a_1, ..., a_k, a \rangle \in A^*$ by ha.

Definition 3.1. A *finite dynamic game with perfect information and ordinal payoffs* is a tuple $\langle A, H, N, \iota, \{\succsim_i\}_{i \in N} \rangle$ whose elements are:

- A finite set of actions A.
- A finite set of histories $H \subseteq A^*$ which is closed under prefixes (that is, if $h \in H$ and $h' \in A^*$ is a prefix of h, then $h' \in H$). The null history $\langle\rangle$, denoted by $\emptyset$, is an element of H and is a prefix of every history. A history $h \in H$ such that, for every $a \in A$, $ha \notin H$, is called a *terminal history*. The set of terminal histories is denoted by Z. $D = H \backslash Z$ denotes the set of non-terminal or *decision* histories. For every decision history $h \in D$, we denote by $A(h)$ the set of actions available at h, that is, $A(h) = \{a \in A : ha \in H\}$.
- A finite set N of players.

[19] As remarked in Footnote 17, both authors use the less general definition of selection function where $f : \Omega \times \mathcal{E} \to \Omega$, that is, for every state ω and event E, there is a unique state closest to ω where E is true.

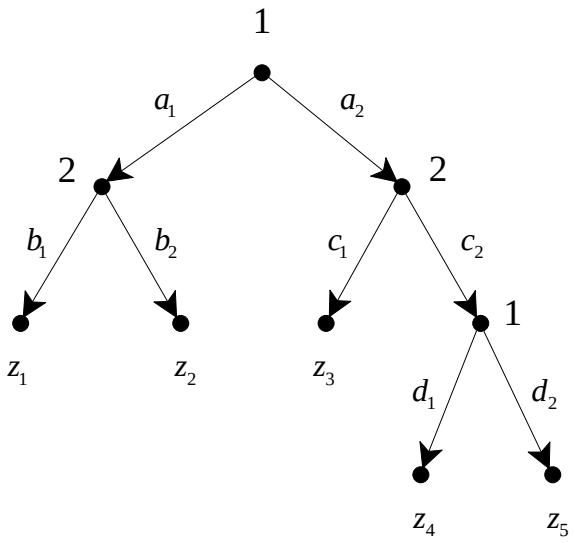

Figure 3: A perfect information game

- A function $\iota : D \to N$ that assigns a player to each decision history. Thus $\iota(h)$ is the player who moves at history h. For every $i \in N$, let $D_i = \iota^{-1}(i)$ be the set of histories assigned to player i.
- For every player $i \in N$, $\succsim_i$ is an ordinal ranking of the set Z of terminal histories.

The ordinal ranking of player i is normally represented by means of an ordinal *utility* (or *payoff*) *function* $U_i : Z \to \mathbb{R}$ satisfying the property that $U_i(z) \geq U_i(z')$ if and only if $z \succsim_i z'$.

Histories will be denoted more succinctly by listing the corresponding actions, without angled brackets and without commas; thus instead of writing $\langle \emptyset, a_1, a_2, a_3, a_4 \rangle$ we simply write $a_1a_2a_3a_4$.

An example of a perfect-information game is shown in Figure 3 in the form of a tree. Each node in the tree represents a history of prior moves and is labeled with the player whose turn it is to move. For example, at history a_2c_2

Player 1 \ Player 2	b_1c_1	b_1c_2	b_2c_1	b_2c_2
a_1d_1	z_1	z_1	z_2	z_2
a_1d_2	z_1	z_1	z_2	z_2
a_2d_1	z_3	z_4	z_3	z_4
a_2d_2	z_3	z_5	z_3	z_5

Figure 4: The strategic form corresponding to the game of Figure 3

it is Player 1's turn to move (after his initial choice of a_2 followed by Player 2's choice of c_2) and he has to choose between two actions: d_1 and d_2. The terminal histories (the leaves of the tree, denoted by z_j, $j = 1, ..., 5$) represent the possible outcomes and each player i is assumed to have a preference relation $\succsim_i$ over the set of terminal histories (in Figure 3 the players' preferences over the terminal histories have been omitted).

In their seminal book, von Neumann and Morgenstern (1944) showed that a dynamic game can be reduced to a normal-form (or strategic-form) game by defining strategies as complete, contingent plans of action. In the case of perfect-information games a strategy for a player is a function that associates with every decision history assigned to that player one of the choices available there. For example, a possible strategy of Player 1 in the game of Figure 3 is (a_1, d_2). A profile of strategies (one for each player) determines a unique path from the null history (the root of the tree) to a terminal history (a leaf of the tree). Figure 4 shows the strategic-form corresponding to the extensive form of Figure 3.

How should a model of a dynamic game be constructed? One approach in

the literature (see, for example, Aumann (1995)) has been to consider models of the corresponding strategic-form (the type of models considered in Section 2: see Definition 2.3). However, there are several conceptual issues that arise in this context. Recall that the interpretation of $s_i = \sigma_i(\omega)$ suggested in Section 2 is that at state ω player i "chooses" strategy s_i. Now consider a model of the game of Figure 3 and a state ω where $\sigma_1(\omega) = (a_1, d_2)$. What does it mean to say that Player 1 "chooses" strategy (a_1, d_2)? The first part of the strategy, namely a_1, can be interpreted as a description of Player 1's actual choice to play a_1, but the second part of the strategy, namely d_2, has no such interpretation: if Player 1 in fact plays a_1 then he knows that he will not have to make any further choices and thus it is not clear what it means for him to "choose" to play d_2 in a situation that is made impossible by his decision to play a_1.[20] Thus it does not seem to make sense to interpret $\sigma_1(\omega) = (a_1, d_2)$ as 'at state ω Player 1 chooses (a_1, d_2)'. Perhaps the correct interpretation is in terms of a more complex sentence such as 'Player 1 chooses to play a_1 and if - contrary to this - he were to play a_2 and Player 2 were to follow with c_2, then Player 1 would play d_2'. Thus while in a simultaneous game the association of a strategy of player i to a state can be interpreted as a description of player i's actual behavior at that state, in the case of dynamic games this interpretation is no longer valid, since one would end up describing not only the actual behavior of player i but also his counterfactual behavior. Methodologically, this is not satisfactory: if it is considered to be necessary to specify what a player would do in situations that do not occur in the state under consideration, then one should model the counterfactual explicitly. But why should it be necessary to specify at state ω (where Player 1 is playing a_1) what he would do at the counterfactual history a_2c_2? Perhaps what matters is not so much what Player 1 would actually do there but what Player 2 believes that Player 1 would do: after all, Player 2 might not know that Player 1 has decided to play a_1 and needs to consider what to do in the eventuality that Player 1 actually ends up playing a_2. So, perhaps, the strategy of Player 1 is to be interpreted as having two components: (1) a description of Player 1's behavior and (2) a conjecture in the mind of Player 2 about what Player 1 would do.[21] If this is the correct interpretation, then one

[20] For this reason, some authors (see, for example, Perea (2012)), instead of using strategies, use the weaker notion of "plan of action" introduced by Rubinstein (1991). A plan of action for a player only contains choices that are not ruled out by his earlier choices. For example, the possible plans of action for Player 1 in the game of Figure 3 are a_1, (a_2, d_1) and (a_2, d_2). However, most of the issues raised below apply also to plans of action. The reason for this is that a choice of player i at a later decision history of his may be counterfactual at a state because of the choices of *other* players (which prevent that history from being reached).

[21] This interpretation of strategies has in fact been put forward in the literature for the case of

could object - from a methodological point of view - that it would be preferable to disentangle the two components and model them explicitly.

In order to clarify these issues it seems that, in the case of dynamic games, one should not adopt the models of Section 2 and instead consider a more general notion of model, where states are described in terms of players' *actual behavior* and any relevant counterfactual propositions are modeled explicitly.

We shall first consider models obtained by adding a counterfactual selection function (see Definition 2.4) to an interactive belief structure (see Definition 2.1) and show that such models are not adequate.

Fix a dynamic game Γ with perfect information and consider the following candidate for a definition of a model of Γ: it is a tuple $\langle N, \Omega, \{\mathcal{B}_i\}_{i\in N}, f, \zeta\rangle$ where $\langle N, \Omega, \{\mathcal{B}_i\}_{i\in N}\rangle$ is an interactive belief structure, $f : \Omega\times\mathcal{E} \to 2^\Omega$ is a counterfactual selection function and $\zeta : \Omega \to Z$ is a function that associates with every state $\omega \in \Omega$ a terminal history (recall that Z denotes the set of terminal histories in Γ).[22] Given a history h in the game, we denote by $[h]$ the set of states where h is reached, that is, $[h] = \{\omega \in \Omega : h \text{ is a prefix of } \zeta(\omega)\}$. We take the set of admissible hypotheses $\mathcal{E}$ (the domain of $f(\omega, \cdot)$) to be the set of propositions of the form "history h is reached", that is, $\mathcal{E} = \{[h] : h \in H\}$ (where H is the set of histories in the game). We now discuss a number of issues that arise in such models.

In the models of Section 2 it was assumed that a player always knows his own strategy (see (4) above). Should a similar assumption be made within the context of dynamic games? That is, suppose that at state ω player i takes action a; should we assume that player i believes that she takes action a? For example, consider a model of the game of Figure 3 in which there are two states, ω and ω', such that $\mathcal{B}_2(\omega) = \{\omega, \omega'\}$ and $\zeta(\omega) = a_1b_1$. Then at state ω Player 2 takes action b_1. Should we require that Player 2 take action b_1 also at ω' (since $\omega' \in \mathcal{B}_2(\omega)$)? The answer is negative: the relation $\mathcal{B}_2$ represents the prior or *initial* beliefs of Player 2 (that is, her beliefs before the game begins) and Player 2 may be uncertain as to whether Player 1 will play a_1 or a_2 and plan to play herself b_1 in the former case and c_1 in the latter case. Thus it makes perfect sense to have $\zeta(\omega') = a_2c_1$. If we want to rule out uncertainty by a player about her action at a decision history of hers, then we need to impose the following restriction:

mixed strategies (which we do not consider in this chapter, given our non-probabilistic approach): see, for example, Aumann and Brandenburger (1995) and the references given there in Footnote 7.

[22] Samet (1996) was the first to propose models of perfect-information games where states are described not in terms of strategies but in terms of terminal histories.

If h is a decision history of player i, a an action available to i at h
and ha a prefix of $\zeta(\omega)$ then, $\forall \omega' \in \mathcal{B}_i(\omega)$, (10)
if h is a prefix of $\zeta(\omega')$ then ha is a prefix of $\zeta(\omega')$.

The above definition can be stated more succinctly in terms of events. If E and F are two events, we denote by $E \rightarrow F$ the event $\neg E \cup F$ (we use the negation symbol $\neg$ to denote the set-theoretic complement, that is, $\neg E$ is the complement of event E). Thus $E \rightarrow F$ captures the material conditional. Recall that, given a history h in the game, $[h] = \{\omega \in \Omega : h \text{ is a prefix of } \zeta(\omega)\}$; recall also that D_i denotes the set of decision histories of player i and $A(h)$ the set of choices available at h. Then (10) can be stated as follows:

$$\begin{aligned} &\forall h \in D_i, \forall a \in A(h), \\ &[ha] \subseteq \mathbb{B}_i([h] \rightarrow [ha]). \end{aligned} \tag{11}$$

In words: if, at a state, player i takes action a at her decision history h, then she believes that if h is reached then she takes action a.[23]

A more subtle issue is whether we should require (perhaps as a condition of rationality) that a player have correct beliefs about what she would do in a situation that she believes will not arise. Consider, for example, the (part of a) model of the game of Figure 3 illustrated in Figure 5. The first line gives $\mathcal{B}_2$, the doxastic accessibility relation of Player 2, the second line the function ζ (which associates with every state a terminal history) and the third line is a partial illustration of the counterfactual selection function: the arrow from state β to state α labeled with the set $\{\alpha, \delta\}$ represents $f(\beta, \{\alpha, \delta\}) = \{\alpha\}$ and the arrow from γ to δ labeled with the set $\{\alpha, \delta\}$ represents $f(\gamma, \{\alpha, \delta\}) = \{\delta\}$.[24] Note that the event that Player 1 plays a_2 is the set of states ω where a_2 is a prefix of $\zeta(\omega)$: $[a_2] = \{\alpha, \delta\}$. Recall that $E \rightrightarrows F$ denotes the counterfactual conditional 'if E were the case then F would be the case'. Now, $[a_2] \rightrightarrows [a_2c_1] = \{\gamma, \delta\}$ and

[23] Note that, if at state ω player i believes that history h will *not* be reached ($\forall \omega' \in \mathcal{B}_i(\omega)$, $\omega' \notin [h]$) then $\mathcal{B}_i(\omega) \subseteq \neg[h] \subseteq [h] \rightarrow [ha]$, so that $\omega \in \mathbb{B}_i([h] \rightarrow [ha])$ and therefore (11) is trivially satisfied (even if $\omega \in [ha]$).

[24] On the other hand, we have not represented the fact that $f(\alpha, \{\alpha, \delta\}) = \{\alpha\}$, which follows from point 3 of Definition 2.4 (since $\alpha \in \{\alpha, \delta\}$) and the fact that $f(\delta, \{\alpha, \delta\}) = \{\delta\}$, which also follows from point 3 of Definition 2.4. We have also omitted other values of the selection function f, which are not relevant for the discussion below.

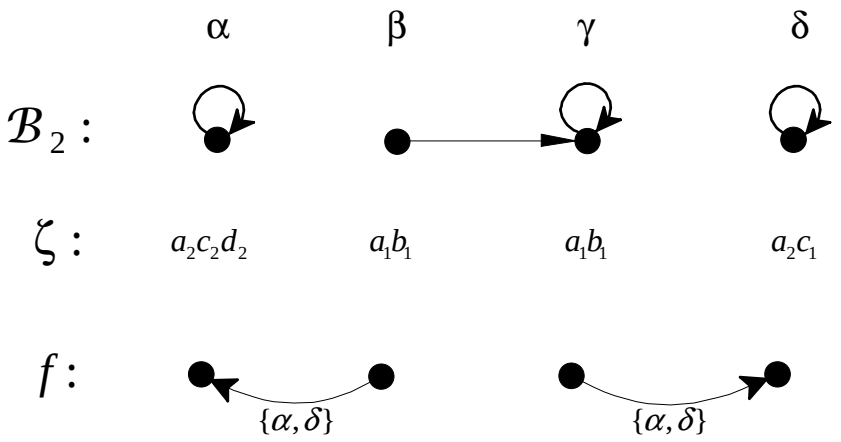

Figure 5: Part of a model of the game of Figure 3

$[a_2] \rightrightarrows [a_2c_2] = \{\alpha, \beta\}$.[25] Thus $\beta \in [a_2] \rightrightarrows [a_2c_2]$ and also $\beta \in \mathbb{B}_2([a_2] \rightrightarrows [a_2c_1])$.[26] That is, at state β it is actually the case that if Player 1 were to play a_2 then Player 2 would respond with c_2, but Player 2 erroneously believes that (if Player 1 were to play a_2) she would respond with c_1.

As a condition of rationality, should one rule out situations like the one illustrated in Figure 5? Shouldn't a rational player have introspective access to what she would do in all the relevant hypothetical situations? In general, it seems that the answer should be negative, since what an individual would do in counterfactual situations may depend on external circumstances (e.g. states of the world or actions of other individuals) which the player might be unaware of (or have erroneous beliefs about). In such circumstances no amount of introspection can aid the individual in acquiring awareness of, or forming correct beliefs about, these external circumstances. This observation might not be applicable to games of complete information, but might be relevant in situations of incomplete information.[27]

There are further issues to be examined. Consider, again, the perfect information game of Figure 3 and a model of this game in which there is a state, say α, where Player 1 plays a_2. Is a_2 a rational choice for Player 1? Answering this question requires answering the following two questions:

Q1. What *will* Player 2 do next?
Q2. What *would* Player 2 do if, instead, a_1 had been chosen?

Let us start with Q1. Consider a model (different from the one described in Figure 5) where at state α the play of the game is $a_2c_2d_1$ (that is, $\zeta(\alpha) = a_2c_2d_1$). If there is "common recognition" of rationality, Player 1 will ask himself how a rational Player 2 will respond to his initial choice of a_2. In order to determine what is rational for Player 2 to do at state α, we need to examine Player 2's beliefs at α. Suppose that Player 2 mistakenly believes that Player 1 will play

[25] Recall that, by Definition 2.4, since $\alpha \in [a_2]$, $f(\alpha, [a_2]) = \{\alpha\}$, so that, since $\alpha \in [a_2c_2]$ (because a_2c_2 is a prefix of $\zeta(\alpha) = a_2c_2d_2$), $\alpha \in [a_2] \rightrightarrows [a_2c_2]$. Furthermore, since $f(\beta, [a_2]) = \{\alpha\}$, $\beta \in [a_2] \rightrightarrows [a_2c_2]$. There is no other state ω where $f(\omega, [a_2]) \subseteq [a_2c_2]$. Thus $[a_2] \rightrightarrows [a_2c_2] = \{\alpha, \beta\}$. The argument for $[a_2] \rightrightarrows [a_2c_1] = \{\gamma, \delta\}$ is similar.

[26] Since $\mathcal{B}_2(\beta) = \{\gamma\}$ and $\gamma \in [a_2] \rightrightarrows [a_2c_1]$, $\beta \in \mathbb{B}_2([a_2] \rightrightarrows [a_2c_1])$. Recall that the material conditional 'if E is the case then F is the case' is captured by the event $\neg E \cup F$, which we denote by $E \to F$. Then $[a_2] \to [a_2c_1] = \{\beta, \gamma, \delta\}$ and $[a_2] \to [a_2c_2] = \{\alpha, \beta, \gamma\}$, so that we also have, trivially, that $\beta \in \mathbb{B}_2([a_2] \to [a_2c_1])$ and $\beta \in \mathbb{B}_2([a_2] \to [a_2c_2])$.

[27] Recall that a game is said to have complete information if the game itself is common knowledge among the players. On the other hand, in a situation of incomplete information at least one player lacks knowledge of some of the aspects of the game, such as the preferences of her opponents, or the actions available to them, or the possible outcomes, etc.

a_1 ($\alpha \in \mathbb{B}_2[a_1]$); for example, $\mathcal{B}_2(\alpha) = \{\beta\}$ and $\beta \in [a_1]$. Furthermore, suppose that $f(\beta, [a_2]) = \{\gamma\}$ and $\gamma \in [a_2c_2d_2]$. Then at α Player 2 believes that if it were the case that Player 1 played a_2 then the play of the game would be $a_2c_2d_2$ ($\alpha \in \mathbb{B}_2([a_2] \rightrightarrows [a_2c_2d_2])$), in particular, she believes that *Player 1 would play* d_2. Since, at state α, Player 1 in fact plays a_2, Player 2 will be surprised: she will be informed that Player 1 played a_2 and that she herself has to choose between c_1 and c_2. What choice she will make depends on her beliefs after she learns that (contrary to her initial expectation) Player 1 played a_2, that is, on her *revised beliefs*. In general, no restrictions can be imposed on Player 2's revised beliefs after a surprise: for example, it seems perfectly plausible to allow Player 2 to become convinced that the play of the game will be $a_2c_2d_1$; in particular, that *Player 1 will play* d_1. The models that we are considering do not provide us with the tools to express such a change of mind for Player 2: if one takes as her revised beliefs her initial beliefs about counterfactual statements that have a_2 as an antecedent, then - since $\alpha \in \mathbb{B}_2([a_2] \rightrightarrows [a_2c_2d_2])$ - one is forced to rule out the possibility that after learning that Player 1 played a_2 Player 2 will believe that the play of the game will be $a_2c_2d_1$. Stalnaker argues that imposing such restrictions is conceptually wrong, since it is based on confounding causal with epistemic counterfactuals:

> "Player 2 has the following initial belief: Player 1 would choose d_2 on his second move [after his initial choice of a_2] if he had a second move. This is a causal 'if' – an 'if' used to express 2's opinion about 1's disposition to act in a situation that she believes will not arise. [...] But to ask what Player 2 would believe about Player 1 if she learned that she was wrong about 1's first choice is to ask a completely different question – this 'if' is epistemic; it concerns Player 2's belief revision policies, and not Player 1's disposition to act." (Stalnaker (1998), p. 48; with small changes to adapt the quote to the game of Figure 3.)

Let us now turn to question Q2. Suppose that, as in the previous example, we are considering a model of the game of Figure 3 and a state α in that model where

$$\alpha \in [a_2c_2d_1] \cap \mathbb{B}_1[a_2] \cap \mathbb{B}_2[a_1b_1] \cap \mathbb{B}_1\mathbb{B}_2[a_1b_1] \tag{12}$$

(for example, (12) is satisfied if $\mathcal{B}_1(\alpha) = \{\alpha\}, \mathcal{B}_2(\alpha) = \{\beta\}$ and $\beta \in [a_1b_1]$). Thus at α Player 1 plays a_2. Is this a rational choice? The answer depends on how Player 2 would respond to the alternative choice of a_1. However, since the

rationality of playing a_2 has to be judged relative to Player 1's beliefs, what matters is not what Player 2 would actually do (at state α) if a_1 were to be played, but what Player 1 believes that Player 2 would do. How should we model such beliefs of Player 1? Again, one possibility is to refer to Player 1's beliefs about counterfactuals with $[a_1]$ as antecedent. If we follow this route, then we restrict the possible beliefs of Player 1; in particular, it cannot be the case that Player 1 believes that if he were to play a_1 then Player 2 would play b_2, that is, we cannot have $\alpha \in \mathbb{B}_1([a_1] \rightrightarrows [a_1b_2])$. Intuitively, the reason is as follows (the formal proof will follow). The counterfactual selection function is meant to capture causal relationships between events. As Stalnaker points out, in the counterfactual world where a player makes a choice different from the one that he is actually making, the prior beliefs of the other players must be the same as in the actual world (by changing his choice he cannot cause the prior beliefs of his opponents to change):

> "I know, for example, that it would be irrational to cooperate in a one-shot prisoners' dilemma because I know that in the counterfactual situation in which I cooperate, my payoff is less than it would be if I defected. And while I have the capacity to influence my payoff (negatively) by making this alternative choice, I could not, by making this choice, influence your prior beliefs about what I will do; that is, your prior beliefs will be the same, in the counterfactual situation in which I make the alternative choice, as they are in the actual situation." (Stalnaker (2006), p. 178)

The formal proof that it cannot be the case that $\alpha \in \mathbb{B}_1([a_1] \rightrightarrows [a_1b_2])$ goes as follows. Suppose that $\alpha \in \mathbb{B}_1([a_1] \rightrightarrows [a_1b_2])$ and fix an arbitrary $\omega \in \mathcal{B}_1(\alpha)$. By (12), since $\alpha \in \mathbb{B}_1[a_2]$, $\omega \in [a_2]$. Fix an arbitrary $\delta \in f(\omega, [a_1])$. Since $\alpha \in \mathbb{B}_1([a_1] \rightrightarrows [a_1b_2])$, and $\omega \in \mathcal{B}_1(\alpha)$, $\omega \in [a_1] \rightrightarrows [a_1b_2]$, that is, $f(\omega, [a_1]) \subseteq [a_1b_2]$. Thus

$$\delta \in [a_1b_2]. \tag{13}$$

Since $\omega \in \mathcal{B}_1(\alpha)$ and $\alpha \in \mathbb{B}_1\mathbb{B}_2([a_1b_1])$, $\omega \in \mathbb{B}_2[a_1b_1]$. By the above remark, at δ the initial beliefs of Player 2 must be the same as at ω.[28] Hence $\delta \in \mathbb{B}_2[a_1b_1]$. By definition, $\delta \in \mathbb{B}_2[a_1b_1]$ if and only if $\mathcal{B}_2(\delta) \subseteq [a_1b_1]$. Thus, since $[a_1b_1] \subseteq \neg[a_1] \cup [a_1b_1] = [a_1] \rightarrow [a_1b_1]$, $\mathcal{B}_2(\delta) \subseteq [a_1] \rightarrow [a_1b_1]$, that is, $\delta \in \mathbb{B}_2([a_1] \rightarrow [a_1b_1])$. Now, (11) requires that, since $\delta \in [a_1b_2]$, $\delta \in \mathbb{B}_2([a_1] \rightarrow [a_1b_2])$. Hence, since

[28] As shown above, at state ω Player 1 chooses a_2; $f(\omega, [a_1])$ is the set of states closest to ω where Player 1 chooses a_1; in these states Player 2's prior beliefs must be the same as at ω, otherwise by switching from a_2 to a_1 Player 1 would cause a change in Player 2's prior beliefs.

$\delta \in \mathbb{B}_2[a_1]$, $\delta \in \mathbb{B}_2[a_1b_2]$, contradicting (13).
In words, since $\alpha \in \mathbb{B}_1\mathbb{B}_2[a_1b_1]$, at every state ω that Player 1 considers possible at α ($\omega \in \mathcal{B}_1(\alpha)$) Player 2 believes that the play of the game is a_1b_1, that is, that she herself will play b_1. If $\alpha \in \mathbb{B}_1([a_1] \rightrightarrows [a_1b_2])$ then $\omega \in [a_1] \rightrightarrows [a_1b_2]$; thus, if δ is a state closest to ω where Player 1 plays a_1, then (by the second property of counterfactual selection functions) at δ Player 2 will actually play b_2. Since Player 1, by changing his choice, cannot cause the initial beliefs of Player 2 to change, Player 2 must have at δ the same beliefs that she has at ω, namely that she will play b_1. Thus at state δ Player 2 believes that she will take action b_1 at her decision history a_1 while in fact she will take action b_2, contradicting the requirement expressed in (11).

Thus we have shown that adding a counterfactual selection function to an interactive belief structure does not provide an adequate notion of model of a dynamic game. The approach followed in the literature[29] has been to do without an "objective" counterfactual selection function f and to introduce in its place "subjective" counterfactual functions f_i (one for each player $i \in N$) representing the players' dispositions to revise their beliefs under various hypotheses.[30] This is the topic of the next section.

4 Belief revision

We will now consider models of dynamic games defined as tuples $\langle N, \Omega, \{\mathcal{B}_i\}_{i\in N}, \{\mathcal{E}_i, f_i\}_{i\in N}, \zeta\rangle$ where - as before - $\langle N, \Omega, \{\mathcal{B}_i\}_{i\in N}\rangle$ is an interactive belief structure and $\zeta : \Omega \to Z$ is a function that associates with every state $\omega \in \Omega$ a terminal history. The new element is $\{\mathcal{E}_i, f_i\}$ (for every player $i \in N$), which is a subjective counterfactual selection function, defined as follows.

Definition 4.1. For every $i \in N$, let $\mathcal{E}_i \subseteq 2^\Omega \backslash \varnothing$ be a set of events representing potential items of information or admissible hypotheses for player i.[31] A *subjective counterfactual selection function* is a function $f_i : \Omega \times \mathcal{E}_i \to 2^\Omega$ that satisfies

[29] See, for example, Arló-Costa and Bicchieri (2007), Baltag et al. (2009), Battigalli et al. (2013), Board (1998), Bonanno (2011), Clausing (2004), Halpern (1999a), Samet (1996).

[30] In Clausing (2004) there is also an objective counterfactual selection function, but it is used only to encode the structure of the game in the syntax.

[31] For example, in a perfect-information game one can take $\mathcal{E}_i = \{[h] : h \in D_i\}$, that is, the set of propositions of the form "decision history h of player i is reached" or $\mathcal{E}_i = \{[h] : h \in H\}$, the set of propositions corresponding to all histories (in which case $\mathcal{E}_i = \mathcal{E}_j$ for any two players i and j).

the following properties: $\forall \omega \in \Omega$, $\forall E, F \in \mathcal{E}_i$,

1. $f_i(\omega, E) \neq \varnothing$,
2. $f_i(\omega, E) \subseteq E$,
3. if $\mathcal{B}_i(\omega) \cap E \neq \varnothing$ then $f_i(\omega, E) = \mathcal{B}_i(\omega) \cap E$,
4. if $E \subseteq F$ and $f_i(\omega, F) \cap E \neq \varnothing$ then $f_i(\omega, E) = f_i(\omega, F) \cap E$.

The event $f_i(\omega, E)$ is interpreted as the set of states that player i would consider possible, at state ω, under the supposition that (or if informed that) E is true. Condition 1 requires these suppositional beliefs to be consistent. Condition 2 requires that E be indeed considered true. Condition 3 says that if E is compatible with the initial beliefs then the suppositional beliefs coincide with the initial beliefs conditioned on event E.[32] Condition 4 is an extension of 3: if E implies F and E is compatible (not with player i's prior beliefs but) with the *posterior* beliefs that she would have if she supposed (or learned) that F were the case (let's call these her posterior F-beliefs), then her beliefs under the supposition (or information) that E must coincide with her posterior F-beliefs conditioned on event E.[33]

Remark 1. *If $\mathcal{E}_i = 2^\Omega \backslash \varnothing$ then Conditions 1-4 in Definition 4.1 imply that, for every $\omega \in \Omega$, there exists a "plausibility" relation Q_i^ω on Ω which is complete ($\forall \omega_1, \omega_2 \in \Omega$, either $\omega_1 Q_i^\omega \omega_2$ or $\omega_2 Q_i^\omega \omega_1$ or both) and transitive ($\forall \omega_1, \omega_2, \omega_3 \in \Omega$, if $\omega_1 Q_i^\omega \omega_2$ and $\omega_2 Q_i^\omega \omega_3$ then $\omega_1 Q_i^\omega \omega_3$) and such that, for every non-empty $E \subseteq \Omega$, $f_i(\omega, E) = \{x \in E : xQ_i^\omega y, \forall y \in E\}$. The interpretation of $\alpha Q_i^\omega \beta$ is that - at state ω and according to player i - state α is at least as plausible as state β. Thus $f_i(\omega, E)$ is the set of most plausible states in E (according to player i at state ω). If $\mathcal{E}_i \neq 2^\Omega \backslash \varnothing$ then Conditions 1-4 in Definition 4.1 are necessary but not sufficient for the existence of such a plausibility relation. The existence of a plausibility relation that rationalizes the function $f_i(\omega, \cdot) : \mathcal{E}_i \to 2^\Omega$ is necessary and sufficient for the belief revision policy encoded in $f_i(\omega, \cdot)$ to be compatible with the theory of belief revision introduced in Alchourrón et al. (1985), known as the AGM theory (see Bonanno (2009)).*

One can associate with each function f_i an operator $\rightrightarrows_i : \mathcal{E}_i \times 2^\Omega \to 2^\Omega$ as follows:

[32] Note that it follows from Condition 3 and seriality of $\mathcal{B}_i$ that, for every $\omega \in \Omega$, $f_i(\omega, \Omega) = \mathcal{B}_i(\omega)$, so that one could simplify the definition of model by dropping the relations $\mathcal{B}_i$ and recovering the initial beliefs from the set $f_i(\omega, \Omega)$. We have chosen not to do so in order to maintain continuity in the exposition.

[33] Although widely accepted, this principle of belief revision is not uncontroversial (see Rabinowicz (1996) and Stalnaker (2009)).

$$E \rightrightarrows_i F = \{\omega \in \Omega : f_i(\omega, E) \subseteq F\}. \tag{14}$$

Possible interpretations of the event $E \rightrightarrows_i F$ are "according to player i, if E were the case, then F would be true" (Halpern (1999a)) or "if informed that E, player i would believe that F" (Stalnaker (1998)) or "under the supposition that E, player i would believe that F" (Arló-Costa and Bicchieri (2007)).[34]

Thus the function f_i can be used to model the full epistemic state of player i; in particular, how player i would revise her prior beliefs if she contemplated information that contradicted those beliefs. However, as pointed out by Stalnaker,

> "It should be noted that even with the addition of the belief revision structure to the epistemic models [...], they remain static models. A model of this kind represents only the agent's beliefs at a fixed time [before the game is played], together with the policies or dispositions to revise her beliefs that she has at that time. The model does not represent any actual revisions that are made when new information is actually received."(Stalnaker (2006), p. 198.)[35]

Condition (11) rules out the possibility that a player may be uncertain about her own choice of action at decision histories of hers that are not ruled out by her initial beliefs. Does a corresponding restriction hold for revised beliefs? That is, suppose that at state ω player i erroneously believes that her decision history h will not be reached ($\omega \in [h]$ but $\omega \in \mathbb{B}_i \neg [h]$); suppose also that a is the action that she will choose at h ($\omega \in [ha]$). Is it necessarily the case that, according to her revised beliefs on the suppositions that h is reached, she believes that she takes action a? That is, is it the case that $\omega \in [h] \rightrightarrows_i [ha]$? In general, the answer is negative. For example, consider a model of the game of Figure 3 in which there are states α, β and γ such that $\alpha \in [a_1 b_1]$, $\mathcal{B}_2(\alpha) = \{\beta\}$, $\beta \in [a_2 c_1]$, $f_2(\alpha, [a_1]) = \{\gamma\}$ and $\gamma \in [a_1 b_2]$. Then we have that at state α Player 2 will in fact take action b_1 (after being surprised by Player 1's choice of a_1) and yet, according to her revised beliefs on the supposition that Player 1 plays a_1, she does not believe

[34] Equivalently, one can think of $\rightrightarrows_i$ as a conditional belief operator $\mathbb{B}_i(\cdot|\cdot)$ with the interpretation of $\mathbb{B}_i(F|E)$ as 'player i believes F given information/supposition E' (see, for example, Board (2004) who uses the notation $\mathbb{B}_i^E(F)$ instead of $\mathbb{B}_i(F|E)$).

[35] The author goes on to say that "The models can be enriched by adding a temporal dimension to represent the dynamics, but doing so requires that the knowledge and belief operators be time indexed..." For a model where the belief operators are indeed time indexed and represent the actual beliefs of the players when actually informed that it is their turn to move, see Bonanno (2013).

that she would take action b_1 (in fact she believes that she would take action b_2): $\alpha \notin [a_1] \rightrightarrows_i [a_1b_1]$. In order to rule this out we need to impose the following strengthening of (11):[36]

$$\begin{aligned} &\forall h \in D_i,\ \forall a \in A(h), \\ &[ha] \subseteq ([h] \rightrightarrows_i [ha]). \end{aligned} \tag{15}$$

Should (15) be considered a necessary component of a definition of rationality? Perhaps so, if the revised beliefs were the actual beliefs of player i when she is actually informed (to her surprise) that her decision history h has been reached. In that case it may be reasonable to assume that - as the player makes up her mind about what to do - she forms correct beliefs about what she is going to do. However, we stressed above that the models we are considering are static models: they represent the initial beliefs and disposition to revise those beliefs at the beginning of the game. Given this interpretation of the revised beliefs as hypothetical beliefs conditional on various suppositions, it seems that violations of (15) might be perfectly rational. To illustrate this point, consider the above example with the following modification: $f_2(\alpha, [a_1]) = \{\alpha, \gamma\}$. It is possible that if Player 1 plays a_1, Player 2 is indifferent between playing b_1 or b_2 (she gets the same payoff). Thus she can coherently form the belief that if - contrary to what she expects - Player 1 were to play a_1, then she might end up choosing either b_1 or b_2: $\alpha \in [a_1] \rightrightarrows_i ([a_1b_1] \cup [a_1b_2])$. Of course, when actually faced with the choice between b_1 and b_2 she will have to break her indifference and pick one action (perhaps by tossing a coin): in the example under consideration (where $\alpha \in [a_1b_1]$) she will pick b_1 (perhaps because the outcome of the coin toss is Heads: something she will know then but cannot know at the beginning).

How can rationality of choice be captured in the models that we are considering? Various definitions of rationality have been suggested in the litera-

[36] (15) is implied by (11) whenever player i's initial beliefs do not rule out h. That is, if $\omega \in \neg\mathbb{B}_i\neg[h]$ (equivalently, $\mathcal{B}_i(\omega) \cap [h] \neq \varnothing$) then, for every $a \in A(h)$,

$$\text{if } \omega \in [ha] \text{ then } \omega \in ([h] \rightrightarrows_i [ha]). \qquad \text{(F1)}$$

In fact, by Condition 3 of Definition 4.1 (since, by hypothesis, $\mathcal{B}_i(\omega) \cap [h] \neq \varnothing$),

$$f_i(\omega, [h]) = \mathcal{B}_i(\omega) \cap [h]. \qquad \text{(F2)}$$

Let $a \in A(h)$ be such that $\omega \in [ha]$. Then, by (11), $\omega \in \mathbb{B}_i([h] \to [ha])$, that is, $\mathcal{B}_i(\omega) \subseteq \neg[h] \cup [ha]$. Thus $\mathcal{B}_i(\omega) \cap [h] \subseteq (\neg[h] \cap [h]) \cup ([ha] \cap [h]) = \varnothing \cup [ha] = [ha]$ (since $[ha] \subseteq [h]$) and therefore, by (F2), $f_i(\omega, [h]) \subseteq [ha]$, that is, $\omega \in [h] \rightrightarrows_i [ha]$.

ture, most notably *material rationality* and *substantive rationality* (Aumann (1995; 1998)). The former notion is weaker in that a player can be found to be irrational only at decision histories of hers that are actually reached. The latter notion, on the other hand, is more stringent since a player can be judged to be irrational at a decision history h of hers even if she knows that h will not be reached. We will focus on the weaker notion of material rationality. We want to define a player's rationality as a proposition, that is, an event. Let $u_i : Z \to \mathbb{R}$ be player i's ordinal utility function (representing her preferences over the set of terminal histories Z) and define $\pi_i : \Omega \to \mathbb{R}$ by $\pi_i(\omega) = u_i(\zeta(\omega))$. For every $x \in \mathbb{R}$, let $[\pi_i \leq x]$ be the event that player i's payoff is not greater than x, that is, $[\pi_i \leq x] = \{\omega \in \Omega : \pi_i(\omega) \leq x\}$ and, similarly, let $[\pi_i > x] = \{\omega \in \Omega : \pi_i(\omega) > x\}$. Then we say that player i is materially rational at a state if, for every decision history h of hers that is actually reached at that state and for every real number x, it is not the case that she believes – under the supposition that h is reached – that (1) her payoff from her actual choice would not be greater than x and (2) her payoff would be greater than x if she were to take an action different from the one that she is actually taking (at that history in that state).[37]

Formally this can be stated as follows (recall that D_i denotes the set of decision histories of player i and $A(h)$ the set of actions available at h):

$$\begin{array}{l} \text{Player } i \text{ is } \textit{materially rational} \text{ at } \omega \in \Omega \text{ if, } \forall h \in D_i, \forall a \in A(h) \\ \text{if } ha \text{ is a prefix of } \zeta(\omega) \text{ then, } \forall b \in A(h),\ \forall x \in \mathbb{R}, \\ ([ha] \rightrightarrows_i [\pi_i \leq x]) \to \neg\,([hb] \rightrightarrows_i [\pi_i > x])\,. \end{array} \tag{16}$$

Note that, in general, we cannot replace the antecedent $[ha] \rightrightarrows_i [\pi_i \leq x]$ with $\mathbb{B}_i([ha] \to [\pi_i \leq x])$, because at state ω player i might initially believe that h will not be reached, in which case it would be trivially true that $\omega \in \mathbb{B}_i([ha] \to [\pi_i \leq x])$. Thus, in general, her rationality is judged on the basis of her *revised* beliefs on the supposition that h is reached. Note, however, that if $\omega \in \neg\mathbb{B}_i\neg[h]$, that is, if at ω she does not rule out the possibility that h will be reached and $a \in A(h)$ is the action that she actually takes at ω ($\omega \in [ha]$), then, for every event F, $\omega \in \mathbb{B}_i([ha] \to F)$ if and only if $\omega \in ([ha] \rightrightarrows_i F)$.[38] Note also that, according

[37] This is a "local " definition in that it only considers, for every decision history of player i, a change in player i's choice at that decision history and not also at later decision histories of hers (if any). One could make the definition of rationality more stringent by simultaneously considering changes in the choices at a decision history and subsequent decision histories of the same player (if any).

[38] Proof. Suppose that $\omega \in [ha] \cap \neg\mathbb{B}_i\neg[h]$. As shown in Footnote 36 (see (F2)),

$$\mathcal{B}_i(\omega) \cap [h] = f_i(\omega, [h]). \qquad \text{(G1)}$$

to (16), a player is trivially rational at any state at which she does not take any actions.

The solution concept which is normally used for perfect-information games is the backward-induction solution, which is obtained as follows. Start from a decision history followed only by terminal histories (such as history a_1a_2 in the game of Figure 6) and pick an action there that is payoff-maximizing for the corresponding player; delete the selected decision history, turn it into a terminal history and associate with it the payoff vector corresponding to the selected choice; repeat the procedure until all the decision histories have been exhausted. For example, the backward-induction solution of the game of Figure 6 selects actions d_3 and d_1 for Player 1 and d_2 for Player 2, so that the corresponding outcome is d_1.

Does initial common belief that all the players are materially rational (according to (16)) imply backward induction in perfect-information games? The answer is negative.[39] To see this, consider the perfect-information game shown in Figure 6 and the model of it shown in Figure 7.[40] First of all, note that the common belief relation $\mathcal{B}^*$ is obtained by adding to $\mathcal{B}_2$ the pair (β, β); thus, in particular, $\mathcal{B}^*(\beta) = \{\beta, \gamma\}$. We want to show that both players are materially rational at both states β and γ, so that at state β it is common belief that both players are materially rational, despite that fact that the play of the game at β

Since $[ha] \subseteq [h]$,

$$\mathcal{B}_i(\omega) \cap [h] \cap [ha] = \mathcal{B}_i(\omega) \cap [ha]. \qquad \text{(G2)}$$

As shown in Footnote 36, $f_i(\omega, [h]) \subseteq [ha]$ and, by Condition 1 of Definition 4.1, $f_i(\omega, [h]) \neq \varnothing$. Thus $f_i(\omega, [h]) \cap [ha] = f_i(\omega, [h]) \neq \varnothing$. Hence, by Condition 4 of Definition 4.1,

$$f_i(\omega, [h]) \cap [ha] = f_i(\omega, [ha]). \qquad \text{(G3)}$$

By intersecting both sides of (G1) with $[ha]$ and using (G2) and (G3) we get that $\mathcal{B}_i(\omega) \cap [ha] = f_i(\omega, [ha])$.

[39] In fact, common belief of material rationality does not even imply a Nash equilibrium outcome. A *Nash equilibrium* is a strategy profile satisfying the property that no player can increase her payoff by unilaterally changing her strategy. A Nash equilibrium *outcome* of a perfect-information game is a terminal history associated with a Nash equilibrium. A backward-induction solution of a perfect-information game can be written as a strategy profile and is always a Nash equilibrium.

[40] In Figure 6, for every terminal history, the top number associated with it is Player 1's utility and the bottom number is Player 2's utility. In Figure 7 we have only represented parts of the functions f_1 and f_2, namely that $f_1(\gamma, \{\alpha, \beta, \delta\}) = \{\delta\}$ and $f_2(\beta, \{\alpha, \beta, \delta\}) = f_2(\gamma, \{\alpha, \beta, \delta\}) = \{\alpha\}$ (note that $[a_1] = \{\alpha, \beta, \delta\}$). Similar examples can be found in Board (2004), Clausing (2004), Rabinowicz (2000), Stalnaker (1998).

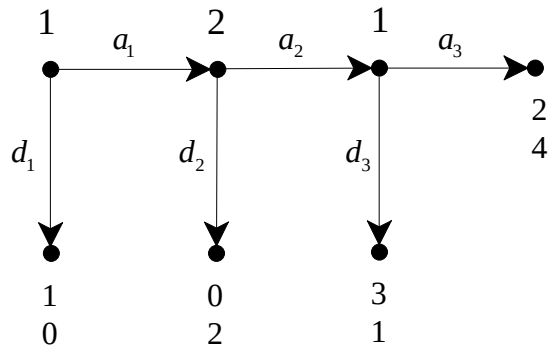

Figure 6: A perfect information game

is $a_1a_2d_3$, while the outcome associated with the backward-induction solution is d_1 (furthermore, there is no Nash equilibrium whose associated outcome is $a_1a_2d_3$). Clearly, Player 1 is materially rational at state β (since he obtains his largest possible payoff); he is also rational at state γ because he knows that he plays d_1, obtaining a payoff of 1, and believes that if he were to play a_1 Player 2 would respond with d_2 and give him a payoff of zero: this belief is encoded in $f_1(\gamma, [a_1]) = \{\delta\}$ (where $[a_1] = \{\alpha, \beta, \delta\}$) and $\zeta(\delta) = a_1d_2$. Player 2 is trivially materially rational at state γ since she does not take any actions there. Now consider state β. Player 2 initially erroneously believes that Player 1 will end the game by playing d_1; however, Player 1 is in fact playing a_1 and thus Player 2 will be surprised. Her initial disposition to revise her beliefs on the supposition that Player 1 plays a_1 is such that she would believe that she herself would play a_2 and Player 1 would follow with a_3, thus giving her the largest possible payoff (this belief is encoded in $f_2(\beta, [a_1]) = \{\alpha\}$ and $\zeta(\alpha) = a_1a_2a_3$). Hence she is rational at state β, according to (16).

In order to obtain the backward-induction solution, one needs to go beyond common initial belief of material rationality. Proposals in the literature include the notions of epistemic independence (Stalnaker (1998)), strong belief (Battigalli and Siniscalchi (2002)), stable belief (Baltag et al. (2009)), substantive rationality (Aumann (1995), Halpern (2001)). For an overview of this literature

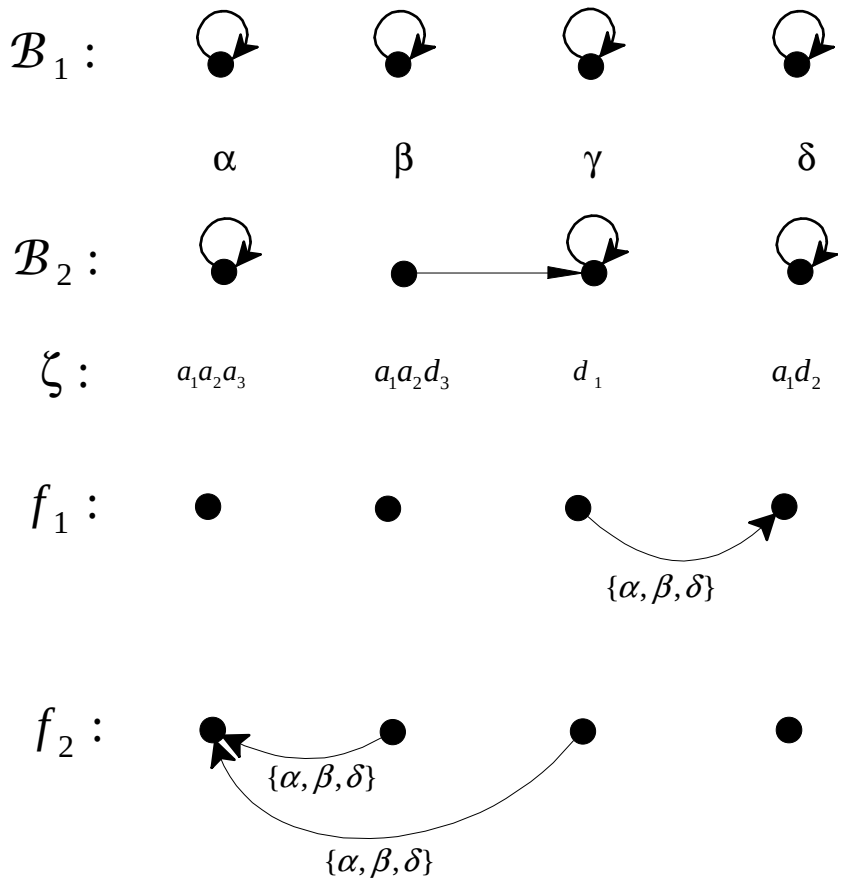

Figure 7: A model of the game of Figure 6

the reader is referred to Brandenburger (2007) and Perea (2007).

It is worth stressing that *in the models considered above, strategies do not play any role*: states are described in terms of the players' actual behavior along a play of the game.[41] One could view a player's strategy as her (conditional) beliefs about what she would do under the supposition that each of her decision histories is reached. However, the models considered so far do not guarantee that a player's revised beliefs select a unique action at each of her decision histories. For example, consider a model of the game of Figure 3 in which there are states α, β and γ such that $\alpha \in [a_2c_1]$, $\mathcal{B}_2(\alpha) = \{\alpha\}$, $\beta \in [a_1b_1]$, $\gamma \in [a_1b_2]$ and $f_2(\alpha, [a_1]) = \{\beta, \gamma\}$. Then, at state α, Player 2 knows that she will take action c_1 and, according to her revised beliefs on the supposition that Player 1 plays a_1, she is uncertain as to whether she would respond to a_1 by playing b_1 or b_2 (perhaps she is indifferent between b_1 and b_2, because she would get the same payoff in either case). One could rule this possibility out by imposing the following restriction:

$$\begin{array}{l}\forall h \in D_i, \forall a, b \in A(h), \forall \omega, \omega', \omega'' \in \Omega, \text{ if } \omega', \omega'' \in f_i(\omega, [h]) \\ \text{and } ha \text{ is a prefix of } \zeta(\omega') \text{ and } hb \text{ is a prefix of } \zeta(\omega'') \text{ then } a = b.\end{array} \tag{17}$$

If (17) is imposed then one can associate with every state a unique strategy for every player. However, as Samet (1996) points out, in this setup strategies would be cognitive constructs rather than objective counterfactuals about what a player would actually do at each of her decision histories.

5 Conclusion

Roughly speaking, a player's choice is rational if, according to what the player believes, there is no other choice which is better for her. Thus, in order to be able to assess the rationality of a player, one needs to be able to represent both the player's choices and her beliefs. The notion of a model of a game does precisely this. We have discussed a number of conceptual issues that arise in attempting to represent not only the actual beliefs but also the counterfactual or hypothetical beliefs of the players. These issues highlight the complexity of defining the notion of rationality in dynamic games and of specifying an

[41] For an example of epistemic models of dynamic games where strategies do play a role see Perea (2014).

appropriate interpretation of the hypothesis that there is "common recognition" of rationality.

A strategy of a player in a dynamic game with perfect information, according to the definition first proposed by von Neumann and Morgenstern (von Neumann and Morgenstern (1944)), is a complete contingent plan specifying a choice of action for every decision history that belongs to that player.[42] We have argued that using the notion of strategy in models of dynamic games is problematic, since it implicitly introduces counterfactual considerations, both objective (in terms of statements about what a player would do in situations that do not arise) and subjective (in terms of the hypothetical or conditional beliefs of the players). Such counterfactuals ought to be modeled explicitly. We first considered the use of objective counterfactuals in models of dynamic games, but concluded that such counterfactuals are inadequate, since they express causal relationships, while it is epistemic counterfactuals that seem to be relevant in terms of evaluating the rationality of choices. We then considered models that make exclusive use of subjective (or epistemic) counterfactuals and showed that in these models strategies do not play any role and can thus be dispensed with.

The models of dynamic games considered above, however, are not the only possibility. Instead of modeling the epistemic states of the players in terms of their prior beliefs and prior dispositions to revise those beliefs in a static framework, one could model the actual beliefs that the players hold at the time at which they make their choices. In such a framework the players' initial belief revision policies (or dispositions to revise their initial beliefs) can be dispensed with: the analysis can be carried out entirely in terms of the actual beliefs at the time of choice. This alternative approach is put forward in Bonanno (2013), where an epistemic characterization of backward induction is provided that does not rely on (objective or subjective) counterfactuals.[43,44]

[42] In general dynamic games, a strategy specifies a choice for every information set of the player.

[43] Bonanno (2013) uses a dynamic framework where the set of "possible worlds" is given by state-instant pairs (ω, t). Each state ω specifies the entire play of the game (that is, a terminal history) and, for every instant t, (ω, t) specifies the history that is reached at that instant (in state ω). A player is said to be active at (ω, t) if the history reached in state ω at date t is a decision history of his. At every state-instant pair (ω, t) the beliefs of the active player provide an answer to the question "what will happen if I take action a?", for every available action a. A player is said to be rational at (ω, t) if either he is not active there or the action he ends up taking at state ω is optimal given his beliefs at (ω, t). Backward induction is characterized in terms of the following event: the first mover (at date 0) (i) is rational and has correct beliefs, (ii) believes that the active player at date 1 is rational and has correct beliefs, (iii) believes that the active player at date 1 believes that the active player at date 2 is rational and has correct beliefs, etc.

[44] The focus of this chapter has been on the issue of modeling the notion of rationality and

A Summary of notation

The following table summarizes the notation used in this chapter.

"common recognition" of rationality in dynamic games with perfect information. Alternatively one can use the AGM theory of belief revision to provide foundations for refinements of Nash equilibrium in dynamic games. This is done in Bonanno (2011; forthcominga) where a notion of perfect Bayesian equilibrium is proposed for general dynamic games (thus allowing for imperfect information). Perfect Bayesian equilibria constitute a refinement of subgame-perfect equilibria and are a superset of sequential equilibria. The notion of sequential equilibrium was introduced by Kreps and Wilson (1982).

Notation	Interpretation
Ω	Set of states.
$\mathcal{B}_i$	Player i's binary "doxastic accessibility" relation on Ω. The interpretation of $\omega\mathcal{B}_i\omega'$ is that at state ω player i considers state ω' possible: see Definition 2.1.
$\mathcal{B}_i(\omega) = \{\omega' \in \Omega : \omega\mathcal{B}_i\omega'\}$	Belief set of player i at state ω.
$\mathbb{B}_i : 2^\Omega \to 2^\Omega$	Belief operator of player i. If $E \subseteq \Omega$ then $\mathbb{B}_i E$ is the set of states where player i believes E, that is, $\mathbb{B}_i E = \{\omega \in \Omega : \mathcal{B}_i(\omega) \subseteq E\}$.
$\mathcal{B}^*$	Common belief relation on the set of states Ω (the transitive closure of the union of the $\mathcal{B}_i$'s).
$\mathbb{B}^* : 2^\Omega \to 2^\Omega$	Common belief operator.
$\langle N, \{S_i, \succsim_i\}_{i\in N}\rangle$	Strategic-form game: see Definition 2.2.
$f : \Omega \times \mathcal{E} \to 2^\Omega$	Objective counterfactual selection function. The event $f(\omega, E)$ is interpreted as "the set of states closest to ω where E is true": see Definition 2.4.
$E \rightrightarrows F = \{\omega \in \Omega : f(\omega, E) \subseteq F\}$	The interpretation of $E \rightrightarrows F$ is "the set of states where it is true that if E were the case then F would be the case."
$\langle A, H, N, \iota, \{\succsim_i\}_{i\in N}\rangle$	Dynamic game with perfect information. See Definition 3.1.
$f_i : \Omega \times \mathcal{E}_i \to 2^\Omega$	Subjective counterfactual selection function. The event $f_i(\omega, E)$ is interpreted as the set of states that player i would consider possible, at state ω, under the supposition that (or if informed that) E is true: see Definition 4.1.
$E \rightrightarrows_i F = \{\omega \in \Omega : f_i(\omega, E) \subseteq F\}$	The event $E \rightrightarrows_i F$ is interpreted as "the set of states where, according to player i, if E were the case, then F would be true".

References

C. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50:510–530, 1985.

H. Arló-Costa and C. Bicchieri. Knowing and supposing in games of perfect information. *Studia Logica*, 86:353–373, 2007.

R. Aumann. What is game theory trying to accomplish? In K. Arrow and S. Honkapohja, editors, *Frontiers in economics*, pages 28–76. Basil Blackwell, Oxford, 1985.

R. Aumann. Backward induction and common knowledge of rationality. *Games and Economic Behavior*, 8:6–19, 1995.

R. Aumann. On the centipede game. *Games and Economic Behavior*, 23:97–105, 1998.

R. Aumann and A. Brandenburger. Epistemic conditions for Nash equilibrium. *Econometrica*, 63:1161–1180, 1995.

A. Baltag, S. Smets, and J. Zvesper. Keep hoping for rationality: a solution to the backward induction paradox. *Synthese*, 169:301–333, 2009.

P. Battigalli and G. Bonanno. Recent results on belief, knowledge and the epistemic foundations of game theory. *Research in Economics*, 53:149–225, 1999.

P. Battigalli and M. Siniscalchi. Strong belief and forward induction reasoning. *Journal of Economic Theory*, 106:356–391, 2002.

P. Battigalli, A. Di-Tillio, and D. Samet. Strategies and interactive beliefs in dynamic games. In D. Acemoglu, M. Arellano, and E. Dekel, editors, *Advances in Economics and Econometrics. Theory and Applications: Tenth World Congress*. Cambridge University Press, Cambridge, 2013.

E. Ben-Porath. Nash equilibrium and backwards induction in perfect information games. *Review of Economic Studies*, 64:23–46, 1997.

D. Bernheim. Rationalizable strategic behavior. *Econometrica*, 52:1002–1028, 1984.

O. Board. Belief revision and rationalizability. In I. Gilboa, editor, *Theoretical aspects of rationality and knowledge (TARK VII)*. Morgan Kaufman, San Francisco, 1998.

O. Board. Dynamic interactive epistemology. *Games and Economic Behavior*, 49: 49–80, 2004.

O. Board. The equivalence of Bayes and causal rationality in games. *Theory and Decision*, 61:1–19, 2006.

G. Bonanno. A syntactic approach to rationality in games with ordinal payoffs. In G. Bonanno, W. van der Hoek, and M. Wooldridge, editors, *Logic and the Foundations of Game and Decision Theory (LOFT 7)*, volume 3 of *Texts in Logic and Games*, pages 59–86. Amsterdam University Press, 2008.

G. Bonanno. Rational choice and *AGM* belief revision. *Artificial Intelligence*, 173: 1194–1203, 2009.

G. Bonanno. *AGM* belief revision in dynamic games. In K. Apt, editor, *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge*, TARK XIII, pages 37–45, New York, 2011. ACM. doi: http://doi.acm.org/10.1145/2000378.2000383.

G. Bonanno. A dynamic epistemic characterization of backward induction without counterfactuals. *Games and Economics Behavior*, 78:31–45, 2013.

G. Bonanno. *AGM*-consistency and perfect Bayesian equilibrium. Part I: definition and properties. *International Journal of Game Theory*, forthcominga. doi: 10.1007/s00182-011-0296-4.

G. Bonanno. Epistemic foundations of game theory. In H. van Ditmarsch, J. Halpern, W. van der Hoek, and B. Kooi, editors, *Handbook of Epistemic Logic*. College Publications, forthcomingb.

G. Bonanno and K. Nehring. Assessing the truth axiom under incomplete information. *Mathematical Social Sciences*, 36:3–29, 1998.

G. Bonanno and K. Nehring. Common belief with the logic of individual belief. *Mathematical Logic Quarterly*, 46:49–52, 2000.

A. Brandenburger. The power of paradox: some recent developments in interactive epistemology. *International Journal of Game Theory*, 35:465–492, 2007.

C. Camerer. *Behavioral game theory: experiments in strategic interaction*. Princeton University Press, Princeton, 2003.

T. Clausing. Doxastic conditions for backward induction. *Theory and Decision*, 54:315–336, 2003.

T. Clausing. Belief revision in games of perfect information. *Economics and Philosophy*, 20:89–115, 2004.

B. de Bruin. *Explaining games: the epistemic programme in game theory*. Springer, 2010.

E. Dekel and F. Gul. Rationality and knowledge in game theory. In D. Kreps and K. Wallis, editors, *Advances in economics and econometrics*, pages 87–172. Cambridge University Press, 1997.

Y. Feinberg. Subjective reasoning - dynamic games. *Games and Economic Behavior*, 52:54–93, 2005.

M. Gerstung, H. Nakhoul, and N. Beerenwinkel. Evolutionary games with affine fitness functions: applications to cancer. *Dynamic Games and Applications*, 1:370–385, 2011.

J. Halpern. Hypothetical knowledge and counterfactual reasoning. *International Journal of Game Theory*, 28:315–330, 1999a.

J. Halpern. Set-theoretic completeness for epistemic and conditional logic. *Annals of Mathematics and Artificial Intelligence*, 26:1–27, 1999b.

J. Halpern. Substantive rationality and backward induction. *Games and Economic Behavior*, 37:425–435, 2001.

D. Kreps and R. Wilson. Sequential equilibrium. *Econometrica*, 50:863–894, 1982.

E. Pacuit. Strategic reasoning in games, this volume. 2014.

D. Pearce. Rationalizable strategic behavior and the problem of perfection. *Econometrica*, 52:1029–1050, 1984.

A. Perea. Epistemic foundations for backward induction: an overview. In J. van Benthem, D. Gabbay, and B. Löwe, editors, *Interactive logic. Proceedings of the 7th Augustus de Morgan Workshop*, volume 1 of *Texts in Logic and Games*, pages 159–193. Amsterdam University Press, 2007.

A. Perea. *Epistemic game theory: reasoning and choice*. Cambridge University Press, Cambridge, 2012.

A. Perea. Finite reasoning procedures for dynamic games, this volume. 2014.

W. Rabinowicz. Stable revision, or is preservation worth preserving? In A. Fuhrmann and H. Rott, editors, *Logic, action and information: essays on logic in philosophy and artificial intelligence*, pages 101–128. de Gruyter, Berlin, 1996.

W. Rabinowicz. Backward induction in games: on an attempt at logical reconstruction. In W. Rabinowicz, editor, *Value and choice: some common themes in decision theory and moral philosophy*, pages 243–256. Lund Philosophy Reports, 2000.

A. Rubinstein. Comments on the interpretation of game theory. *Econometrica*, 59:909–924, 1991.

D. Samet. Hypothetical knowledge and games with perfect information. *Games and Economic Behavior*, 17:230–251, 1996.

Y. Shoham and K. Leyton-Brown. *Multiagent systems: algorithmic, game-theoretic, and logical foundations*. Cambridge University Press, 2008.

J. M. Smith. *Evolution and the theory of games*. Cambridge University Press, 1982.

R. Stalnaker. A theory of conditionals. In N. Rescher, editor, *Studies in logical theory*, pages 98–112. Blackwell, 1968.

R. Stalnaker. Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, 12:133–163, 1996.

R. Stalnaker. Belief revision in games: forward and backward induction. *Mathematical Social Sciences*, 36:31–56, 1998.

R. Stalnaker. Extensive and strategic forms: games and models for games. *Research in Economics*, 53:293 –319, 1999.

R. Stalnaker. On logics of knowledge and belief. *Philosophical Studies*, 128: 169–199, 2006.

R. Stalnaker. Iterated belief revision. *Erkenntnis*, 128:189–209, 2009.

K. Suzumura. *Rational choice, collective decisions and social welfare*. Cambridge University Press, Cambridge, 1983.

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

J. von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 1944.

E. Zambrano. Counterfactual reasoning and common knowledge of rationality in normal form games. *Topics in Theoretical Economics*, 4:Article 8, 2004.