Counterfactuals and the Prisoner's Dilemma

Giacomo Bonanno*

Department of Economics
University of California
Davis, CA 95616 – USA
e-mail: gfbonanno@ucdavis.edu
URL: http://www.econ.ucdavis.edu/faculty/bonanno/

1. Introduction

In 2011 Harold Camping, president of Family Radio (a California-based Christian radio station), predicted that Rapture (the taking up into heaven of God's elect people) would take place on May 21, 2011. In light of this prediction some of his followers gave up their jobs, sold their homes and spent large sums promoting Camping's claims.[1] Did these people act rationally? Consider also the following hypothetical scenarios. Early in 2012, on the basis of a popular reading of the Mayan calendar, Ann came to believe that the world would end on December 21, 2012. She dropped out of college, withdrew all the money she had in her bank account and decided to spend it all on travelling and enjoying herself. Was her decision rational? Bob smokes two packets of cigarettes a day; when asked if he would still smoke if he knew that he was going to get lung cancer from smoking, he answers 'No'; when asked if he is worried about getting lung cancer, he says that he is not and explains that his grandfather was a heavy smoker all his life and died – cancer free – at the age of 98. Bob believes that, like his grandfather, he is immune from lung cancer. Is Bob's decision to continue smoking rational?

I will argue below that the above questions are closely related to the issue, hotly debated in the literature, whether it can be rational for the players to choose "Cooperation" in the Prisoner's Dilemma game, shown in Figure 1. It is a two-player, simultaneous game where each player has two strategies: "Cooperation" (denoted by *C*) and "Defection" (denoted by *D*). In each cell of the table, the first number is the utility (or payoff) of Player 1 and the second number is the utility of Player 2.

		Player 2	
		C	*D*
Player 1	*C*	2 , 2	0 , 3
	D	3 , 0	1 , 1

Figure 1

The Prisoner's Dilemma game

What constitutes a rational choice for a player? We take the following to be the basic definition of rationality (BDR):

A choice is rational if it is optimal given the
decision-maker's preferences and beliefs. (*BDR*)

More precisely, we say that it is rational for the decision-maker to choose action *a* if there is no other feasible action *b* which − *according to her beliefs* − would yield an outcome that she prefers to the outcome that − again, according to her beliefs − would be a consequence of taking action *a*. According to this definition, the followers of Harold Camping did act rationally when they decided to sell everything and devote themselves to promoting Camping's claim: they believed that the world was soon coming to an end and, presumably, they viewed their proselytizing as "qualifying them for Rapture", undoubtedly an outcome that they preferred to the alternative of enduring the wrath of Judgment Day. Similarly, Ann's decision to live it up in anticipation of the end of the world predicted by the Mayan calendar qualifies as rational, as does Bob's decision to carry on smoking on the belief that − like his grandfather − he will be immune from lung cancer. Thus anybody who argues that the above decisions are *not* rational must be appealing to a stronger definition of rationality than *BDR*: one that denies the rationality of holding those beliefs.

When the rationality of beliefs is called into question, an asymmetry is introduced between preferences and beliefs. Concerning preferences it is a generally accepted principle that *de gustibus non est disputandum* (in matters of taste, there can be no disputes). According to this principle, there is no such thing as an irrational preference. As Rubinstein (2012, p. 49) notes,

> "According to the assumption of rationality in economics, the decision maker is guided by his preferences. But the assumption does not impose a limitation on the reasonableness of preferences. The preferences can be even in direct

> contrast with what common sense might define as the decision maker's interests."

For example, I cannot be judged to be irrational if I prefer an immediate benefit (e.g. from taking a drug) with known negative future consequences (e.g. from addiction) over an immediate sacrifice (e.g. by enduring pain) followed by better long-term health.[2]

In the matter of beliefs, on the other hand, it is generally thought that one *can* contend that some particular beliefs are "unreasonable" or "irrational", by appealing to such arguments as the lack of supporting evidence, the incorrect processing of relevant information, the denial of laws of Nature, etc.

Consider now the following statement by Player 1 in the Prisoner's Dilemma ('COR' stands for 'correlation'):

> "I believe that if I play *C* then Player 2 will play *C* and that if I play *D* then Player 2 will play *D*. Thus, if I play *C* my payoff will be 2 and if I play *D* my payoff will be 1. Hence I have decided to play *C*." (COR$_1$)

Given the reported beliefs, Player 1's decision to play *C* is rational according to definition *BDR*. Thus, in order to maintain that it is not rational, one has to argue that the beliefs expressed in COR$_1$ violate some principle of rationality. In the literature, there are those who claim that Player 1's reported beliefs are irrational and those who claim that those beliefs can be rationally justified, for example by appealing to the symmetry of the game (see, for example, Brams, 1975, and Davis, 1977, 1985) or to special circumstances, such as the players being identical in some sense (e.g. they are identical twins): this has become known as the "Identicality Assumption" (this expression is used, for example, in Bicchieri and Green, 1999, and Gilboa, 1999).

In order to elucidate what is involved in Player 1's belief "if I play *C* then Player 2 will play *C*, and if I play *D* then Player 2 will *D*" we need to address the issue of the role of beliefs and conditionals in game-theoretic reasoning. In Section 2 we discuss the notion of model of a game, which provides an explicit representation of beliefs and choices. After arguing that such models do not allow for an explicit discussion of rational choice, we turn in Sections 3-5 to enriched models that contain an explicit representation of subjunctive conditionals and discuss two alternative approaches: one based on belief revision and the other on objective counterfactuals. In Section 6 we review the few contributions in the literature that have offered a definition of rationality in strategic-form games based on an explicit appeal to counterfactuals. In Section 7 we discuss alternative ways of dealing with the conditionals involved in deliberation and Section 8 concludes.

2. Models of games: beliefs and choices

It is a widely held opinion that the notion of rationality involves the use of counterfactual reasoning. For example, Aumann (1995, p. 15) 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. [I]n interactive decision making – games – you must consider what other people would do if you did something different from what you actually do."

How is counterfactual reasoning incorporated in the analysis of games? The definition of strategic-form game provides only a partial description of an interactive situation. A *game in strategic form with ordinal preferences* is defined as a quintuple $G = \left\langle N, \{S_i\}_{i\in N}, O, z, \{\succsim_i\}_{i\in N} \right\rangle$, where $N = \{1,...,n\}$ is a set of *players*, S_i is the set of *strategies* of (or possible choices for) player $i \in N$, O is a set of possible *outcomes*, $z : S \to O$ is a function that associates an outcome with every strategy profile $s = (s_1,...,s_n) \in S = S_1 \times ... \times S_n$ and $\succsim_i$ is a complete and transitive binary relation on O representing player *i*'s ranking of the outcomes (the interpretation of $o \succsim_i o'$ is that player *i* considers outcome *o* to be at least as good as outcome o').[3] Games are typically represented in *reduced form* by replacing the triple $\left\langle O, z, \{\succsim_i\}_{i\in N} \right\rangle$ with a set of *payoff functions* $\{\pi_i\}_{i\in N}$, where $\pi_i : S \to \mathbb{R}$ is any numerical function that satisfies the property that, for all $s, s' \in S$, $\pi_i(s) \geq \pi_i(s')$ if and only if $z(s) \succsim_i z(s')$, that is, if player *i* considers the outcome associated with *s* to be at least as good as the outcome associated with s'. In the following we will adopt this more succinct representation of strategic-form games (as we did in Figure 1).[4] Thus the definition of strategic-form game only specifies what choices each player has available and how the player ranks the possible outcomes; it is silent on what the player believes. In order to complete the description one needs to introduce the notion of *model of a game*.

Definition 1. Given a strategic-form game *G*, a *model of G* is a triple $\left\langle \Omega, \{\sigma_i\}_{i\in N}, \{\mathcal{B}_i\}_{i\in N} \right\rangle$ where Ω is a set of *states* and, for every player $i \in N$, $\sigma_i : \Omega \to S_i$ is a function that associates with every state $\omega \in \Omega$ a strategy $\sigma_i(\omega) \in S_i$ of player *i* and $\mathcal{B}_i \subseteq \Omega \times \Omega$ is a binary 'doxastic' relation representing the beliefs of player *i*. The interpretation of $\omega \mathcal{B}_i \omega'$ is that at state ω player *i* considers state ω' possible. Let $\mathcal{B}_i(\omega) = \{\omega' \in \Omega : \omega \mathcal{B}_i \omega'\}$; thus $\mathcal{B}_i(\omega)$ is the set of states that player *i* considers possible at state ω.[5]

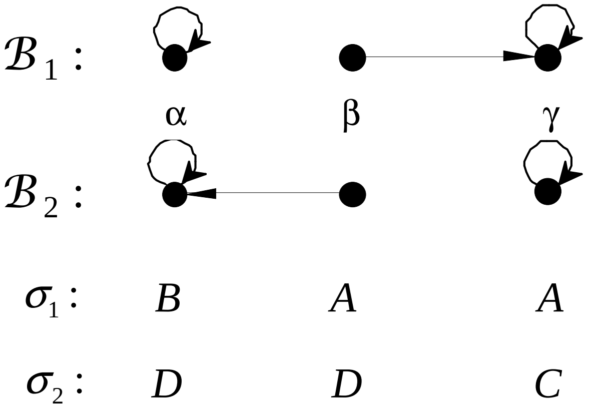
The functions $\{\sigma_i : \Omega \to S_i\}_{i\in N}$ give content to the players' beliefs. If $\sigma_i(\omega) = x \in S_i$ then the usual interpretation is that at state ω player *i* "chooses" strategy *x*. The exact meaning of 'choosing' is not elaborated further in the literature: does it mean that player *i* *has actually played* *x* or that she *will play* *x* or that *x* is the *output of her deliberation process*? We will adopt the latter interpretation: 'player *i* chooses *x*' will be taken to mean 'player *i* has irrevocably made up her mind to play *x*'.

Subsets of Ω are called *events*. Given a state $\omega \in \Omega$ and an event $E \subseteq \Omega$, we say that *at* ω *player i believes E* if and only if $\mathcal{B}_i(\omega) \subseteq E$.

Part *a* of Figure 2 shows a strategic-form game and Part *b* a model of it (we represent a relation $\mathcal{B}$ graphically as follows: $\omega \mathcal{B} \omega'$ – or, equivalently, $\omega' \in \mathcal{B}(\omega)$ – if and only if there is an arrow from ω to ω').

		Player 2	
		C	*D*
Player 1	*A*	2 , 3	0 , 0
	B	0 , 0	3 , 2

(*a*) A strategic-form game



(*b*) A model of the game

Figure 2

State β in the model of Figure 2 represents the following situation: Player 1 has made up her mind to play *A* and Player 2 has made up his mind to play *D*; Player 1 erroneously believes that Player 2 has made up his mind to play *C* ($\mathcal{B}_1(\beta)=\{\gamma\}$ and $\sigma_2(\gamma)=C$) and Player 2 erroneously believes that Player 1 has made up her mind to play *B* ($\mathcal{B}_2(\beta)=\{\alpha\}$ and $\sigma_1(\alpha)=B$).

Remark 1. The model of Figure 2 reflects a standard assumption in the literature, namely that *a player is never uncertain about her own choice*: any uncertainty has to do with the other players' choices. This requirement is expressed formally as follows: for every $\omega'\in\mathcal{B}_i(\omega)$, $\sigma_i(\omega')=\sigma_i(\omega)$ $-$ that is, if at state ω player *i* chooses strategy $x\in S_i$ ($\sigma_i(\omega)=x$) then at ω she believes that she chooses *x*. We shall revisit this point in Section 7.

Returning to the model of Part *b* of Figure 2, a natural question to ask is whether the players are rational at state β. Consider Player 1: according to her beliefs, the outcome is going to be the one associated with the strategy pair (*A*,*C*), with a corresponding payoff of 2 for her. In order to determine whether the decision to play *A* is rational, Player 1 needs to ask herself the question "what would happen if, instead of playing *A*, I were to play *B*?". The model is silent about such counterfactual scenarios. Thus the definition of model introduced above appears to lack the resources to address the issue of rational choice.[6] Before we discuss how to enrich the definition of model (Sections 4 and 5), we turn, in the next section, to a brief digression on the notion of counterfactual.

3. Stalnaker-Lewis selection functions

There are different types of conditionals. A conditional of the form "If John received my message he will be here soon" is called an *indicative* conditional. Conditionals of the form "If I were to drop this vase, it would break" and "If we had not missed the connection, we would be at home now" are called *subjunctive* conditionals; the latter is also an example of a *counterfactual*, namely a conditional with a false antecedent (we did in fact miss the connection). It is controversial how best to classify conditionals and we will not address this issue here. We are interested in the use of conditionals in the analysis of games and thus the relevant conditionals are those that pertain to deliberation.

In the decision-theoretic and game-theoretic literature the conditionals involved in deliberation are usually called “counterfactuals”, as illustrated in the quotation from Aumann (1995) in the previous section and in the following:

> “[R]ational decision-making involves conditional propositions: when a person weighs a major decision, it is rational for him to ask, for each act he considers, what would happen if he performed that act. It is rational, then, for him to consider propositions of the form ‘If I were to do *a*, then *c* would happen’. Such a proposition we shall call a counterfactual.” (Gibbard and Harper, 1978, p. 153.)

With the exception of Shin (1992), Bicchieri and Green (1999), Zambrano (2004) and Board (2006) (whose contributions are discussed in Section 6), the issue of counterfactual reasoning in strategic-form games has not been dealt with explicitly in the literature.[7]

We denote by $\phi > \psi$ the conditional “if ϕ were the case then ψ would be the case”. In the Stalnaker-Lewis theory of conditionals (Stalnaker, 1968, Lewis, 1973) the formula $\phi > \psi$ has a truth value which is determined as follows: $\phi > \psi$ is true at a state ω if and only if ψ is true at all the ϕ-states that are closest (that is, most similar) to ω (a state ω' is a ϕ-state if and only if ϕ is true at ω'). While Stalnaker postulates that, for every state ω and formula ϕ, there is a unique

ϕ-state ω' that is closest to ω, Lewis allows for the possibility that there may be several such states.

The semantic representation of conditionals is done by means of a *selection function* $f : \Omega \times 2^{\Omega} \to 2^{\Omega}$ (where 2^{Ω} denotes the set of subsets of Ω) that associates with every state ω and subset $E \subseteq \Omega$ (representing a proposition) a subset $f(\omega, E) \subseteq E$ interpreted as the states in E that are closest to ω. Several restrictions are imposed on the selection function, but we will skip the details.[8]

Just as the notion of doxastic relation enables us to represent a player’s beliefs without, in general, imposing any restrictions on the content of those beliefs, the notion of selection function enables us to incorporate subjunctive conditionals into a model without imposing any constraints on what ϕ-states ought to be considered most similar to a state where ϕ is not true. A comic strip shows the following dialogue between father and son:[9]

> Father: No, you can’t go.
> Son: But all my friends …
> Father: If all your friends jumped off a bridge, would you jump too?
> Son: Oh, Jeez… Probably.
> Father: What!? Why!?
> Son: Because all my friends did. Think about it: which scenario is more likely? Every single friend I know − many of them levelheaded and afraid of heights − abruptly went crazy at exactly the same time …or the bridge is on fire?

The issue of determining what state(s) ought to be deemed closest to a given state is not a straightforward one. Usually “closeness” is interpreted in terms of a *ceteris paribus* (other things being equal) condition. However, typically *some* background conditions *must* be changed in order to evaluate a counterfactual. Consider, for example, the situation represented by state β in the model of Figure 2. What would be − in an appropriately enriched model − the closest state to β − call it η − where Player 1 plays *B* rather than *A*? It has been argued (we will return to this point later) that it ought to be postulated that η is a state where Player 1 has the same beliefs about Player 2’s choice as in state β. Thus η would be a state where Player 1 plays *B* while believing that Player 2 plays *C*; hence at state η one of the background conditions that describe state β no longer holds, namely, that Player 1 is rational and believes herself to be rational. Alternatively, if one wants to hold this condition constant, then one must postulate that at η Player 1 believes (or at least considers it possible) that Player 2 plays *D* and thus one must change another background condition at β, namely her beliefs about Player 2. We will return to these issue in Section 6.

There is also another issue that needs to be addressed. The selection function *f* is usually interpreted as capturing the notion of “causality” or “objective possibility”. For example, suppose that Ann is facing two faucets, one labeled ‘hot’ and the other ‘cold’, and she needs hot water. Suppose also that the faucets are mislabeled and Ann is unaware of this. Then it would be objectively or causally true that “if Ann turned on the faucet labeled ‘cold’ she would get *hot* water”; however, she could not be judged to be irrational if she expressed the belief “if I turned on the faucet labeled ‘cold’ I would get *cold* water” (and acted on this belief by turning on the faucet labeled ‘hot’). Since what we are interested in is the issue of rational choice, objective counterfactuals do not seem to be the relevant objects to consider: *what matters is not what would in fact be the case but what the agent believes would be the case*. We shall call such beliefs *subjective counterfactuals*. How should these subjective counterfactuals be modeled? There are two options, examined in the following sections.

4. Subjective counterfactuals as dispositional belief revision

One construal of subjective counterfactuals is in terms of a *subjective selection function* $f_i : \Omega \times 2^\Omega \to 2^\Omega$ such that, for every $\omega \in \Omega$ and $E \subseteq \Omega$, $f_i(\omega, E) \subseteq E$. The function f_i is interpreted as expressing, at every state, player *i*'s *initial beliefs together with her disposition to revise those belief under various suppositions*. Fix a state $\omega \in \Omega$ and consider the function $f_{i,\omega} : 2^\Omega \to 2^\Omega$ given by $f_{i,\omega}(E) = f_i(\omega, E)$, for every $E \subseteq \Omega$. This function gives the initial beliefs of player *i* at state ω (represented by the set $f_{i,\omega}(\Omega)$) as well as the set of states that player *i* would consider possible, at state ω, under the supposition that event $E \subseteq \Omega$ is true (represented by the set $f_{i,\omega}(E)$), for every event *E*. Subjective selection functions − with the implied dispositional belief revision policy − have been used extensively in the literature on dynamic games,[10] but (to the best of my knowledge) have not been used in the analysis of strategic-form games, with the exception of Shin (1992) and Zambrano (2004), whose contributions are discussed in Section 6.

In this context, an enriched model of a strategic-form game *G* is a quadruple $\left\langle \Omega, \{\sigma_i\}_{i\in N}, \{\mathcal{B}_i\}_{i\in N}, \{f_i\}_{i\in N} \right\rangle$, where $\left\langle \Omega, \{\sigma_i\}_{i\in N}, \{\mathcal{B}_i\}_{i\in N} \right\rangle$ is as defined in Definition 1 and, for every player *i*, $f_i : \Omega \times 2^\Omega \to 2^\Omega$ is a subjective selection function satisfying the property that, for every state ω, $f_i(\omega, \Omega) = \mathcal{B}_i(\omega)$.[11] Such enriched models would be able to capture the following reasoning of Player 1 in the Prisoner's Dilemma (essentially a restatement of COR$_1$):

> "I have chosen to play *C* and I believe that Player 2 has chosen to play *C* and thus I believe that my payoff will be 2; furthermore, I am happy with my choice of *C* because − under the supposition that I play *D* − I believe that Player 2 would play *D* and thus my payoff would be 1." (COR$_2$)

These beliefs are illustrated by state α in the following enriched model of the Prisoner's Dilemma game of Figure 1: $\Omega = \{\alpha, \beta\}$, $\mathcal{B}_1(\alpha) = \{\alpha\}$, $\mathcal{B}_1(\beta) = \{\beta\}$, $f_1(\alpha, \{\alpha\}) = f_1(\alpha, \Omega) = \{\alpha\}$,

$f_1(\beta,\{\beta\}) = f_1(\beta,\Omega) = \{\beta\}$, $f_1(\alpha,\{\beta\}) = \{\beta\}$, $f_1(\beta,\{\alpha\}) = \{\alpha\}$, $\sigma_1(\alpha) = C$, $\sigma_1(\beta) = D$, $\sigma_2(\alpha) = C$, $\sigma_2(\beta) = D$ (we have omitted the beliefs of Player 2). At state α Player 1 believes that she is playing *C* and Player 2 is playing *C* ($\mathcal{B}_1(\alpha) = \{\alpha\}$ and $\sigma_1(\alpha) = C$ and $\sigma_2(\alpha) = C$); furthermore the proposition "Player 1 plays *D*" is represented by the event $\{\beta\}$ (β is the only state where Player 1 plays *D*) and thus, since $f_1(\alpha,\{\beta\}) = \{\beta\}$ and $\sigma_2(\beta) = D$, Player 1 believes that $-$ under the supposition that she plays *D* $-$Player 2 plays *D* and thus her own payoff would be 1.

Are the beliefs expressed in COR$_2$ compatible with rationality? The principles of "rational" belief revision, that are captured by the properties listed in Footnote 11, are principles of logical coherence of dispositional beliefs[12] and, in general, do not impose any constraints on the content of a counterfactual belief. Thus the above beliefs of Player 1 *could* be rational beliefs, in the sense that they do not violate logical principles or principles of coherence. Those who claim that the beliefs expressed in COR$_2$ are irrational appeal to the argument that they imply a belief by Player 1 that her "switching" from *C* to *D* *causes* Player 2 to change her decision from *C* to *D,* while such a causal effect is ruled out by the fact that each player is making her choice in ignorance of the choice made by the other player (the choices are made "simultaneously"). For example, Harper (1988, p. 25) claims that "a causal independence assumption is part of the idealization built into the normal form" and Stalnaker (1996, p. 138) writes "[I]n a strategic form game, the assumption is that the strategies are chosen independently, which means that the choices made by one player cannot influence the beliefs or the actions of the other players". One can express this point of view by imposing the following restriction on beliefs:

> In an enriched model of a game, if at state ω player *i* considers it possible that his opponent is choosing any one of the strategies $w_1,...,w_m$, then the following must be true for every strategy *x* of player *i*: under the supposition that she plays *x,* player *i* continues to consider it possible that his opponent is choosing any one of the strategies $w_1,...,w_m$ and no other strategies.

This condition can be expressed more succinctly as follows. Given a state ω and a player *i* we denote by $\sigma_{-i}(\omega) = \left(\sigma_1(\omega),...,\sigma_{i-1}(\omega),\sigma_{i+1}(\omega),...,\sigma_n(\omega)\right)$ the strategies chosen at ω by the players other than *i*; furthermore, for every strategy *x* of player *i*, let [*x*] denote the event that (that is, the set of states at which) player *i* plays *x*. Then the above restriction on beliefs can be written as follows (‘IND’ stands for ‘independence’ and ‘subj’ for ‘subjective’)

For every state ω and for every $x \in S_i$,

$$\bigcup_{\omega' \in \mathcal{B}_i(\omega)} \{\sigma_{-i}(\omega')\} = \bigcup_{\omega' \in f_i(\omega,[x])} \{\sigma_{-i}(\omega')\} \qquad (IND_1^{subj})$$

The beliefs expressed in COR_2 violate condition IND_1^{subj}.

Should IND_1^{subj} be viewed as a necessary condition for rational beliefs? This question will be addressed in Section 8.

5. Subjective counterfactuals as beliefs about causality

The usual argument in support of the thesis that, for the Prisoner’s Dilemma, Player 1’s reasoning expressed in COR_1 is fallacious is that even if (e.g. because of symmetry or because of the “identicality assumption”) one agrees that the outcome must be one of the two on the diagonal, the off-diagonal outcomes are nevertheless causally possible. Thus one must distinguish between *causal* (or objective) possibility and *doxastic* (or subjective) possibility and in the process of rational decision making one has to consider the relevant causal possibilities, even if they are ruled out as doxastically impossible. This is where objective counterfactuals become relevant. This line of reasoning is at the core of causal decision theory.[13]

According to this point of view, subjective counterfactuals should be interpreted in terms of the composition of a belief relation $\mathcal{B}_i$ with an objective counterfactual selection function $f : \Omega \times 2^\Omega \to 2^\Omega$. In this approach, an enriched model of a strategic-form game *G* is a quadruple $\left\langle \Omega, \{\sigma_i\}_{i\in N}, \{\mathcal{B}_i\}_{i\in N}, f \right\rangle$, where $\left\langle \Omega, \{\sigma_i\}_{i\in N}, \{\mathcal{B}_i\}_{i\in N} \right\rangle$ is as defined in Definition 1 and $f : \Omega \times 2^\Omega \to 2^\Omega$ is an objective selection function. In this context, $f(\omega, E)$ is the set of states in *E* that would be “causally true” (or objectively true) at state ω if *E* were the case, while

$\bigcup_{\omega' \in \mathcal{B}_i(\omega)} f(\omega', E)$ is the set of states in E that $-$ according to player i's beliefs at state $\omega -$ could be "causally true" if E were the case.

As noted in the previous section, from the point of view of judging the rationality of a choice, what matters is not the "true" causal effect of that choice but what the agent *believes* to be the causal effect of her choice, as illustrated in the example of Section 2 concerning the mislabeled faucets. As another example, consider the case of a player who believes to be engaged $-$as Player 1 $-$ in a Prisoner's Dilemma game, while in fact Player 2 is a computer that will receive as input Player 1's choice and has been programmed to mirror that choice. In this case, in terms of objective counterfactuals, there is perfect correlation between the choices of the two players, so that the best choice of Player 1 would be to play *C*. However, Player 1 may rationally play *D* if she believes that (1) Player 2 will play *D* and (2) if she were to play *C* then Player 2 would still play *D*.

Causal independence, at a state ω, between the choice of player *i* and the choices of her opponents would be expressed by the following restriction on the objective selection function $\big[$recall that $\sigma_{-i}(\omega) = \left(\sigma_1(\omega), ..., \sigma_{i-1}(\omega), \sigma_{i+1}(\omega), ..., \sigma_n(\omega)\right)$ is the profile of strategies chosen at ω by *i*'s opponents and that, for $x \in S_i$, $[x]$ denotes the event that $-$ that is, the set of states where $-$ player *i* chooses strategy *x*; 'obj' stands for 'objective'$\big]$:

For every strategy x of player *i*, if $\omega' \in f(\omega, [x])$, then $\sigma_{-i}(\omega') = \sigma_{-i}(\omega)$. $\quad (IND^{obj})$

However, as noted above, what matters is not whether IND^{obj} holds at state ω but whether player *i believes* that IND^{obj} holds. Hence the following, subjective, version of independence is the relevant condition:

For every strategy x of player i and for every $\omega' \in \mathcal{B}_i(\omega)$, if $\omega'' \in f(\omega', [x])$ then $\sigma_{-i}(\omega'') = \sigma_{-i}(\omega')$. $\quad (IND_2^{subj})$

It is straightforward to check that condition IND_2^{subj} implies condition IND_1^{subj} if one defines $f_i(\omega, E) = \bigcup_{\omega' \in \mathcal{B}_i(\omega)} f(\omega', E)$, for every event *E*; indeed a slightly weaker version of IND_2^{subj} is equivalent to IND_1^{subj}.[14]

We conclude that, since a player may hold erroneous beliefs about the causal effects of her own choices and what matters for rational choice is what the player believes rather than what is "objectively true", there is no relevant conceptual difference between the objective approach discussed in this section and the subjective approach discussed in the previous section.[15]

6. Rationality of choice: discussion of the literature

We are yet to provide a precise definition of rationality in strategic-form games. With the few exceptions described below, there has been no formal discussion of the role of counterfactuals in the analysis of strategic-form games. Aumann (1987) was the first to use the notion of epistemic[16] model of a strategic-form game. His definition of rationality, which we will state in terms of beliefs (rather than the more restrictive notion of knowledge) and call Aumann-rationality, is as follows. Recall that, given a state ω in a model of a game and a player *i*, $\sigma_i(\omega)$ denotes the strategy chosen by player *i* at state ω, while the profile of strategies chosen by the other players is denoted by $\sigma_{-i}(\omega) = \left(\sigma_1(\omega), ..., \sigma_{i-1}(\omega), \sigma_{i+1}(\omega), ..., \sigma_n(\omega)\right)$.

Definition 2. Consider a model of a strategic form game (see Definition 1), a state ω and a player *i*. Player *i*'s choice at state ω is *Aumann-rational* if there is no other strategy s_i of player *i* such that $\pi_i\left(s_i, \sigma_{-i}(\omega')\right) > \pi_i\left(\sigma_i(\omega), \sigma_{-i}(\omega')\right)$ for every $\omega' \in \mathcal{B}_i(\omega)$.[17] That is, player *i*'s choice is rational if it is not the case that player *i* believes that another strategy of hers is strictly better than the chosen strategy.

The above definition is weaker than the definition used in Aumann (1987), since − for simplicity −we have restricted attention to ordinal payoffs and qualitative (that is, non-probabilistic, beliefs).[18] However, *the essential feature of this definition is that it evaluates counterfactual strategies of player i keeping the beliefs of player i constant*. Hence implicit in this definition of rationality is a either a theory of subjective counterfactuals that assumes condition IND_1^{subj} or an objective theory of counterfactuals that assumes condition IND_2^{subj}.

The only attempts (that I am aware of) to bring the relevant counterfactuals to the surface are Shin (1992), Bicchieri and Green (1999), Zambrano (2004) and Board (2006).

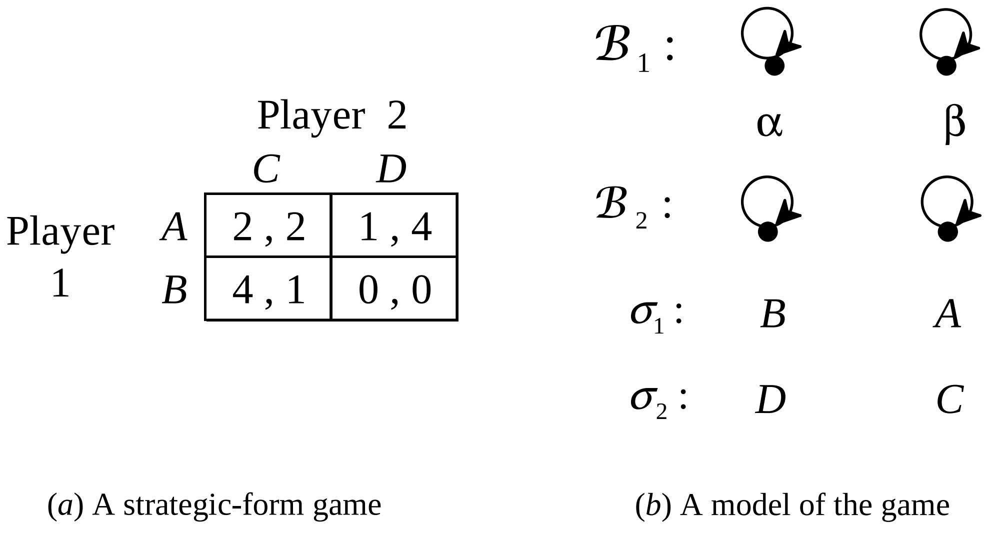
Shin (1992) develops a framework which is very similar to one based on subjective selection functions (as described in Section 4). For each player *i* in a strategic-form game Shin defines a "subjective state space" Ω_i. A point in this space specifies a belief of player *i* about his own choice and the choices of the other players. Such belief assigns probability 1 to player *i*'s own choice (that is, player *i* is assumed to know his own choice). Shin then defines a metric on this space as follows. Let ω be a state where player *i* attaches probability 1 to his own choice, call it *A*, and has beliefs represented by a probability distribution *P* on the strategies of his opponents; the closest state to ω where player *i* chooses a different strategy, say *B*, is a state ω' where player *i* attaches probability 1 to *B* and has the same probability distribution *P* over the strategies of his opponents that he has at ω. This metric allows player *i* to evaluate the counterfactual "if I chose *B* then my payoff would be *x*". Thus Shin imposes *as an axiom* the requirement that player *i* should hold the same beliefs about the other players' choices when contemplating a "deviation" from his actual choice. This assumption corresponds to requirement IND_1^{subj}. Not surprisingly, his main result is that a player is rational with respect to this metric if and only if she is Aumann-rational.

Zambrano's (2004) approach is a mixture of objective and subjective counterfactuals. His analysis is restricted to two-player strategic-form games. First of all, he defines a subjective selection function for player *i*, $f_i : \Omega \times S_i \to \Omega$, which follows Stalnaker (1968) in assuming that, for every hypothesis and every state ω, there is a *unique* world closest to ω where that hypothesis is satisfied; furthermore, the hypotheses consist of the possible strategies of player *i* (the set of strategies S_i), rather than events. He interprets $f_i(\omega, s_i) = \omega'$ as follows: "state ω' is the state closest to ω, according to player *i*, in which player *i* deviates from the strategy prescribed by ω and, instead, plays s_i" (p. 5). He then imposes the requirement that "player *i* is the *only* one that deviates from $\sigma(\omega)$ in $f_i(\omega, s_i)$, that is, $\sigma_j(f_i(\omega, s_i)) = \sigma_j(\omega)$" (Condition F2, p. 5; *j* denotes the other player). This appears to be in the spirit of the objective causal independence assumption

IND_2^{obj}. However, Zambrano does not make use of this requirement, because he focuses on the beliefs of player *i* at the state $f_i(\omega, s_i)$ and uses *these* beliefs to evaluate *both* the original strategy $\sigma_i(\omega)$ *and* the new strategy s_i. He introduces the following definition of rationality:

> "player *i* is W-rational [at state ω] if there is no deviation $s_i \neq \sigma_i(\omega)$ such that strategy s_i is preferred to $\sigma_i(\omega)$ given the belief that player *i* holds *at the state closest to* ω *in which i deviates to* s_i. The interpretation is that the rationality of choosing strategy $\sigma_i(\omega)$ at state ω against a deviation $s_i \neq \sigma_i(\omega)$ is determined with respect to beliefs that arise at the closest state to ω in which s_i is actually chosen, that is, with respect to beliefs at $f_i(\omega, s_i)$." (Zambrano, 2004, p. 6).

Expressed in terms of our qualitative approach, player *i* is W-rational at state ω if there is no strategy s_i of player *i* such that $\pi_i\left(s_i, \sigma_{-i}(\omega')\right) > \pi_i\left(\sigma_i(\omega), \sigma_{-i}(\omega')\right)$ for every $\omega' \in \mathcal{B}_i(f_i(\omega, s_i))$. Hence, unlike Aumann-rationality (Definition 2), the quantification is over $\mathcal{B}_i(f_i(\omega, s_i))$ rather than over $\mathcal{B}_i(\omega)$.[19] *The definition of W-rationality thus disregards the beliefs of player i at state* ω *and focuses instead on the beliefs that player i would have if she changed her strategy*. Since, in general, those hypothetical beliefs can be different from the initial beliefs at state ω, there is no connection between W-rationality and Aumann-rationality. For example, consider the game shown in Part *a* of Figure 3 and the model shown in Part *b*.



		Player 2	
		C	*D*
Player 1	*A*	2 , 2	1 , 4
	B	4 , 1	0 , 0

(*a*) A strategic-form game (*b*) A model of the game

Figure 3

Let the subjective selection function of Player 1 be given by $f_1(\alpha, B) = f_1(\beta, B) = \alpha$ and $f_1(\alpha, A) = f_1(\beta, A) = \beta$. Consider state α where the play is (*B*,*D*) and both players get a payoff of 0. Player 1 is W-rational at state α (where she chooses *B* and believes that Player 2 chooses *D*) because if she were to play *A* (state β) then she would believe that Player 2 played *C* and − given these beliefs − playing *B* is better than playing *A*. However, Player 1 is not Aumann-rational at state α, because the notion of Aumann rationality uses the beliefs of Player 1 at state α to compare *A* to *B* (while the notion of W-rationality uses the beliefs at state β).

Zambrano then shows (indirectly, through the implications of common knowledge of rationality) that W-rationality coincides with Aumann-rationality if one adds the following restriction to the subjective selection function f_i: for every state ω and every strategy $s_i \in S_i$, at the closest state to ω where player *i* plays strategy s_i, the beliefs of player *i* concerning the strategy chosen by the other player (player *j*) are the same as at state ω.[20] This is in the spirit of condition IND_1^{subj}.

Board (2006) uses objective counterfactuals as defined by Stalnaker (1968) (for every hypothesis and every state ω, there is a *unique* world closest to ω where that hypothesis is satisfied). Like Zambrano, Board takes as possible hypotheses the individual strategies of the players: he introduces an *objective* selection function $f : \Omega \times \bigcup_{i \in N} S_i \to \Omega$, that specifies −for every state ω, every player *i* and every strategy $s_i \in S_i$ of player *i* − the unique world $f(\omega, s_i) \in \Omega$ closest to ω where player *i* chooses s_i. Recall that $\sigma_i(\omega)$ denotes the strategy chosen by player *i* at state ω. In accordance with Stalnaker's theory of counterfactuals, Board assumes that $f(\omega, \sigma_i(\omega)) = \omega$, that is, the closest state to ω where player *i* chooses the strategy that he chooses at ω is ω itself. On the other hand, if $s_i \neq \sigma_i(\omega)$ and $f(\omega, s_i) = \omega'$ then it is necessarily the case that $\omega' \neq \omega$, since it must be that $\sigma_i(\omega') = s_i$. What does player *i* believe at state ω about the choices of the other players? As before, let $\mathcal{B}_i$ be the belief relation of player *i* and $\mathcal{B}_i(\omega) = \{\omega' \in \Omega : \omega \mathcal{B}_i \omega'\}$ the belief set of player *i* at state ω. We denote by $S_{-i} = S_1 \times ... \times S_{i-1} \times S_{i+1} \times ... \times S_n$ the set of strategy profiles for the players other than *i*. Then the set of strategy profiles of the opponents that player *i* considers possible at state ω, if she plays her chosen strategy $\sigma_i(\omega)$, is $\bigcup\limits_{\omega' \in \mathcal{B}_i(\omega)} \{\sigma_{-i}(\omega')\}$ = $\{s_{-i} \in S_{-i} : s_{-i} = \sigma_{-i}(\omega') \text{ for some } \omega' \in \mathcal{B}_i(\omega)\}$. On the other hand, what are her beliefs – at state ω – about the strategy profiles of her opponents if she were to choose a strategy $s_i \neq \sigma_i(\omega)$? For every state ω' that she deems possible at state ω (that is, for every $\omega' \in \mathcal{B}_i(\omega)$) she considers the closest state to ω' where she plays s_i, namely $f(\omega', s_i)$, and looks at the choices made by her opponents at state $f(\omega', s_i)$.[21] Thus the set of strategy profiles of the opponents that player *i* would consider possible at state ω, if she were to play a strategy $s_i \neq \sigma_i(\omega)$, is $\bigcup\limits_{\omega' \in \mathcal{B}_i(\omega)} \{\sigma_{-i}(f(\omega', s_i))\}$.[22] Note that, in general, there is no relationship between the sets $\bigcup\limits_{\omega' \in \mathcal{B}_i(\omega)} \{\sigma_{-i}(f(\omega', s_i))\}$ and $\bigcup\limits_{\omega' \in \mathcal{B}_i(\omega)} \{\sigma_{-i}(\omega')\}$; indeed, these two sets might even be disjoint.

Board defines player *i* to be *causally rational* at state ω (where she chooses strategy $\sigma_i(\omega)$) if it is not the case that she believes, at state ω, that there is another strategy $s_i \in S_i$ which would yield a higher payoff than $\sigma_i(\omega)$. His definition is expressed in terms of expected payoff maximization.[23] Since, in general, the two sets $\bigcup\limits_{\omega' \in \mathcal{B}_i(\omega)} \{\sigma_{-i}(f(\omega', s_i))\}$ and $\bigcup\limits_{\omega' \in \mathcal{B}_i(\omega)} \{\sigma_{-i}(\omega')\}$ might be disjoint, causal rationality is consistent with each player choosing Cooperation in the Prisoner’s Dilemma. To see this, consider the following partial model of the Prisoner’s Dilemma game of Figure 1, where, for the sake of brevity, we only specify the beliefs of Player 1 and the objective counterfactuals concerning the strategies of Player 1 and, furthermore, in order to avoid ambiguity, we denote the strategies of Player 1 by *C* and *D* and the strategies of Player 2 by *c* and *d*: $\Omega = \{\alpha, \beta, \gamma, \delta\}$, $\sigma_1(\alpha) = \sigma_1(\beta) = C$, $\sigma_1(\gamma) = \sigma_1(\delta) = D$, $\sigma_2(\alpha) = \sigma_2(\beta) = \sigma_2(\delta) = c$, $\sigma_2(\gamma) = d$, $\mathcal{B}_1(\alpha) = \mathcal{B}_1(\beta) = \{\beta\}$, $\mathcal{B}_1(\gamma) = \{\gamma\}$, $\mathcal{B}_1(\delta) = \{\delta\}$, $f(\alpha, C) = f(\gamma, C) = f(\delta, C) = \alpha$, $f(\beta, C) = \beta$, $f(\alpha, D) = \delta$, $f(\beta, D) = f(\gamma, D) = \gamma$ and $f(\delta, D) = \delta$. Then at state α Player 1 is causally rational: she chooses *C* and believes that her payoff will be 2 (because she believes that Player 2 has chosen *c*: $\mathcal{B}_1(\alpha) = \{\beta\}$ and $\sigma_2(\beta) = c$) and she also believes that if she were to play *D* then Player 2 would play *d* ($\mathcal{B}_1(\alpha) = \{\beta\}$, $f(\beta, D) = \gamma$ and $\sigma_2(\beta) = d$) and thus her payoff would be 1. Note that at state α

Player 1 has incorrect beliefs about what would happen if she played *D*: since $f(\alpha, D) = \delta$ and $\sigma_2(\delta) = c$, the "objective truth" is that if Player 1 were to play *D* then Player 2 would still play *c*, however Player 1 believes that Player 2 would play *d*. Note state α in this model provides a formal representation of the reasoning expressed in COR$_1$.

Board's main result is that a necessary and sufficient condition for causal rationality to coincide with Aumann rationality is the IND_2^{subj} condition of Section 5.[24]

Bicchieri and Green's (1999) aim is to clarify the implications of the "identicality assumption" in the Prisoner's Dilemma game. They enrich the definition of a model of a game (Definition 1) by adding a binary relation $C \subseteq \Omega \times \Omega$ of "nomic accessibility", interpreting $\omega C \omega'$ as "ω' is causally possible relative to ω" in the sense that "everything that occurs at ω' is consistent with the laws of nature that hold at ω" (p. 180). After discussing at length the difference between doxastic possibility (represented by the relations $\mathcal{B}_i$, $i \in N$) and causal possibility (in the spirit of causal decision theory), they raise the question whether it is possible to construe a situation in which it is causally necessary that the choices of the two players in the Prisoner's Dilemma are the same, while their actions are nonetheless causally independent. They suggest that the answer is positive: one could construct an agentive analogue of the Einstein-Podolsky-Rosen phenomenon in quantum mechanics (p. 184). They conclude that there may indeed be a coherent nomic interpretation of the identicality assumption, but such interpretation may be controversial.

In the next section we discuss the issue of whether subjunctive conditionals or counterfactuals – as captured by (subjective or objective) selection functions – are indeed a necessary, or even desirable, tool for the analysis of rational choice.

7. Conditionals of deliberation and pre-choice beliefs

A common feature of all the epistemic/doxastic models of games used in the literature is the assumption that if a player chooses a particular action at state ω then she knows, at state ω, that she chooses that action. This approach thus requires the use of either objective or subjective counterfactuals in order to represent a player's beliefs about the consequences of taking alternative actions. However, several authors have maintained that *it is the essence of* *deliberation that one cannot reason towards a choice if one already knows what that choice will be*. For instance, Shackle (1958, p. 21) remarks that if an agent could predict the option he will choose, his decision problem would be "empty", Ginet (1962, p. 50) claims that "it is conceptually impossible for a person to know what a decision of his is going to be before he makes it", Goldman (1970, p. 194) writes that "deliberation implies some doubt as to whether the act will be done", Spohn (1977, p. 114) states the principle that "any adequate quantitative decision model must not explicitly or implicitly contain any subjective probabilities for acts" (and later [Spohn, 2012, p. 109] writes that "the decision model must not impute to the agent any cognitive or doxastic assessment of his own actions"), Levi (1986, p. 65) states that "the deliberating agent cannot, before choice, predict how he will choose" and coins the phrase "deliberation crowds out prediction" (Levi,1997, p. 81).[25]

Deliberation involves reasoning along the following lines: "if I take action *a*, then the outcome will be *x* and if I take action *b*, then the outcome will be *y*". Indeed it has been argued (DeRose, 2010) that the appropriate conditionals for deliberation are *indicative* conditionals, rather than subjunctive conditional. If I say "if I had left the office at 4 pm I would not have been stuck in traffic", I convey the information that − as a matter of fact − I did not leave the office at 4 pm and thus I am uttering a counterfactual conditional, namely one which has a false antecedent (such a statement would not make sense if uttered before 4 pm). On the other hand, if I say "if I leave the office at 4 pm I will not be stuck in traffic" I am uttering what is normally called an indicative conditional and I am conveying the information that I am evaluating the consequences of a possible future action (such a statement would not make sense if uttered after 4 pm). Concerning the latter conditional, is there a difference between the indicative mood and the subjunctive mood? If I said (before 4 pm) "if I were to leave the office at 4 pm I would not be stuck in traffic", would I be conveying the same information as with the previous indicative conditional? On this point there does not seem to be a clear consensus in the literature. I agree with DeRose's claim that the subjunctive mood conveys different information relative to the indicative mood: its role is to

> "call attention to the possibility that the antecedent is (or will be) false, where one reason one might have for calling attention to the possibility that the antecedent is (or will be) false is that it is quite likely that it is (or will be) false." (DeRose, 2010, p. 10.)

The indicative conditional signals that the decision whether to leave the office at 4 pm is still "open", while the subjunctive conditional intimates that the speaker is somehow ruling out that option: for example, he has made a tentative or firm decision not to leave at 4 pm.

In light of the above discussion it would be desirable to model a player's *deliberation-stage* (or *pre-choice*) beliefs, where the player considers the consequences of all her actions, *without predicting her subsequent decision*. If a state encodes the player's actual choice, then that choice can be judged to be rational or irrational by relating it to the player's pre-choice beliefs. Hence, if one follows this approach, it becomes possible for a player to have the same beliefs in two different states, ω and ω', and be labeled as rational at state ω and irrational at state ω', because the action she ends up taking at state ω is optimal given those beliefs, while the action she ends up taking at state ω' is not optimal given those same beliefs.

A potential objection to this view arises in dynamic games where a player chooses more than once along a given play of the game. Consider a situation where at time t_1 player *i* faces a choice and knows that she might be called upon to make a second choice at a later time t_2. The view outlined above requires player *i* to have "open" beliefs about her choice at time t_1 but also allows her to have beliefs about what choice she will make at the later time t_2. Is this problematic? Several authors have maintained that there is no inconsistency between the principle that one should not attribute to a player beliefs about her current choice and the claim that, on the other hand, one can attribute to the player beliefs about her later choices. For example, Gilboa writes:

> "[W]e are generally happier with a model in which one cannot be said to have beliefs about (let alone knowledge of) one's own choice while making this choice . [O]ne may legitimately ask: Can you truly claim you have no beliefs about your own future choices? Can you honestly contend you do not believe – or even know – that you will not choose to jump out of the window? [T]he answer to these questions is probably a resounding "No". But the emphasis should be on timing: when one considers one's choice tomorrow, one may indeed be quite sure that one will not decide to jump out of the window. However, a future decision should actually be viewed as a decision by a different "agent" of the same decision maker. [...] It is only at the time of choice, within an "atom of decision", that we wish to preclude beliefs about it." (Gilboa,1999, pp. 171 –172)

In a similar vein, Levi (1997 , p. 81) writes that “agent X may coherently assign unconditional credal probabilities to hypotheses as to what he will do when some future opportunity for choice arises. Such probability judgments can have no meaningful role, however, when the opportunity of choice becomes the current one.” Similarly, Spohn (1999, pp. 44 –45) maintains that in the case of sequential decision making, the decision maker can ascribe subjective probabilities to his future − but not to his present − actions. We share the point of view expressed by these authors. If a player moves sequentially at times t_1 and t_2, with $t_1 < t_2$, then at time t_1 she has full control over her immediate choices (those available at t_1) but not over her later choices (those available at t_2). The agent can predict – or form an intention about – her future behavior at time t_2, but she cannot irrevocably decide it at time t_1, just as she can predict – but not decide – how other individuals will behave after her current choice.

Doxastic models of games incorporating deliberation-stage beliefs were recently introduced in Bonanno (2013*b*, 2013*c*) for the analysis of dynamic games. These models allow for a definition of rational choice that is free of (subjective or objective) counterfactuals. Space limitations prevent us from going into the details of these models.

8. Conclusion

Deliberation requires the evaluation of alternatives different from the chosen one: in Aumann's words (1995, p. 15), "you must consider what other people will do if you did something different from what you actually do". Such evaluation thus requires the use of counterfactuals. With very few exceptions (discussed in Section 6), counterfactuals have not been used explicitly in the analysis of rational decision-making in strategic-form games. We argued that objective counterfactuals are not the relevant object to focus on, since in − order to evaluate the rationality of a choice − what matters is not what would in fact be the case but what the player believes would be the case (as illustrated in the example of the mislabeled faucets in Section 3). Hence one should consider subjective counterfactuals. In Sections 4 and 5 we discussed two different ways of modeling subjective counterfactuals, one based on dispositional belief revision and the other on beliefs about causal possibilities and we argued that − for the analysis of strategic-form games (and the Prisoner's Dilemma in particular) − the two approaches are essentially equivalent. We identified a restriction on beliefs (condition IND_1^{subj} of Section 4 and the essentially equivalent condition IND_2^{subj} of Section 5) which in the literature has been taken, either explicitly or implicitly, to be part of a definition of rationality. This restriction requires a player not to change her beliefs about the choices of the other players when contemplating alternative actions to the chosen one. It is a restriction that has been invoked by those who claim that "Cooperation" in the Prisoner's Dilemma cannot be a rational choice (Player 1's beliefs in the Prisoner's Dilemma expressed in COR1 [Section 1] violate it). What motivates this restriction is the view that to believe otherwise is to fail to recognize that the independence of the players' decisions in a strategic-form game makes it causally impossible to affect a change in the opponent's choice merely by "changing one's own choice".

Is this necessarily true? In other words, are there compelling conceptual reasons why IND_1^{subj} (or the essentially equivalent IND_2^{subj}) should be viewed as a necessary condition for rational beliefs? Some authors have claimed that the answer should be negative.

Bicchieri and Green (1999) point out a scenario (an agentive analogue of the Einstein-Podolsky-Rosen phenomenon in quantum mechanics) where causal independence is compatible with correlation and thus it would be possible for a player to coherently believe (a) that her choice is causally independent of the opponent's choice and also (b) that there is correlation between her choice and the opponent's choice, such as the correlation expressed in COR_1.

In a series of contributions, Spohn (2003, 2007, 2010, 2012) put forward a new solution concept, called "dependency equilibrium", which allows for correlation between the players' choices. An example of a dependency equilibrium is (*C,C*) in the Prisoner's Dilemma. Spohn stresses the fact that the notion of dependency equilibrium is consistent with the causal independence of the players' actions:

> "The point then is to conceive the decision situations of the players as somehow jointly caused and as entangled in a dependency equilibrium… [B]y no means are the players assumed to believe in a causal loop between their actions; rather, they are assumed to believe in the possible entanglement as providing a common cause of their actions." (Spohn, 2007, p. 787.)

It should also be pointed out that this "common cause" justification for beliefs is generally accepted when it comes to judging a player's beliefs about the strategies of her opponents: it is a widely held opinion that it can be fully rational for, say, Player 3 to believe – in a simultaneous game − (a) that the choices of Player 1 and Player 2 are causally independent and yet (b) that "if Player 1 plays *x* then Player 2 will play *x* and if Player 1 plays *y* then Player 2 will play *y*". For example, Aumann (1987, p. 16) writes:

> "In a game with more than two players, correlation may express the fact that what 3, say, thinks that 1 will do may depend on what he thinks 2 will do. This has no connection with any overt or even covert collusion between 1 and 2; they may be acting entirely independently. Thus it may be common knowledge that both 1 and 2 went to business school, or perhaps to the same business school; but 3 may not know what is taught there. In that case 3 would think it quite likely that they would take similar actions, without being able to guess what those actions might be."

Similarly, Brandenburger and Friedenberg (2008, p. 32) write that this correlation in the mind of Player 3 between the action of Player 1 and the action of Player 2 "is really just an adaptation to game theory of the usual idea of common-cause correlation."

Thus Player 1's beliefs expressed in COR_1 might perhaps be criticized for being implausible or farfetched, but are not necessarily irrational.

References

Alchourrón, Carlos, Gändenfors, Peter and Makinson, David 1985. "On the logic of theory change: partial meet contraction and revision functions", *The Journal of Symbolic Logic*, 50:510−530.

Arló-Costa, Horacio and Bicchieri, Cristina 2007. "Knowing and supposing in games of perfect information", *Studia Logica* , 86:353−373.

Aumann, Robert 1987. "Correlated equilibrium as an expression of Bayesian rationality", *Econometrica*, 55:1−19.

Aumann, Robert 1995. "Backward induction and common knowledge of rationality," *Games and Economic Behavior*, 8:6−1 9.

Battigalli, Pierpaolo and Bonanno, Giacomo 1999. "Recent results on belief, knowledge and the epistemic foundations of game theory", *Research in Economics*, 53:149−225.

Battigalli, Pierpaolo, Di Tillio, Alfredo and Samet Dov 2013. "Strategies and interactive beliefs in dynamic games", in D. Acemoglu, M. Arellano, E. Dekel (eds.), *Advances in Economics and Econometrics: Theory and Applications: Tenth World Congress*, Cambridge University Press.

Bicchieri, Cristina and Green, Mitchell 1999. “Symmetry arguments for cooperation in the Prisoner’s Dilemma” in C. Bicchieri, R. Jeffrey and B. Skyrms (eds.), *The logic of strategy*, Oxford University Press, pp. 175–195.

Binmore, Ken 2011. *Rational decisions,* Princeton University Press.

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

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

Bonanno, Giacomo 2011. “AGM belief revision in dynamic games”, in: Krzysztof R. Apt (ed.), *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge* (TARK XIII), ACM, New York, pp. 37-45

Bonanno, Giacomo 2013*a*. “Reasoning about strategies and rational play in dynamic games” in J. van Benthem, S. Ghosh and R. Verbrugge (eds.), *Modeling strategic reasoning,* Texts in Logic and Games, Springer, forthcoming.

Bonanno, Giacomo 2013*b*. “A dynamic epistemic characterization of backward induction without counterfactuals”, *Games and Economic Behavior*, 78:31–43.

Bonanno, Giacomo 2013*c*. “An epistemic characterization of generalized backward induction”, Working Paper No. 134, University of California Davis. (http://ideas.repec.org/p/cda/wpaper/13-2.html)

Brams, Steven 1975. “Newcomb’s Problem and Prisoners’ Dilemma”, *The Journal of Conflict Resolution*, 19:496–612.

Brandenburger, Adam and Friedenberg, Amanda 2008. “Intrinsic correlation in games”, *Journal of Economic Theory*, 141:28–67.

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

Davis, Lawrence 1977. “Prisoners, paradox and rationality”, *American Philosophical Quarterly*, 14:319–327.

Davis, Lawrence 1985. “Is the symmetry argument valid?” in R. Campbell and L. Snowden (eds.), *Paradoxes of rationality and cooperation*, University of British Columbia Press, pp. 255–262.

DeRose, Keith 2010. “The conditionals of deliberation”, *Mind*, 119:1–42.

Gibbard, Allan and Harper, William 1978. “Counterfactuals and two kinds of expected utility” in W. Harper, R. Stalnaker and G. Pearce (eds.), *Ifs: conditionals, belief, decision, chance, and time*, D. Reidel, pp. 153–190.

Gilboa, Itzhak 1999. “Can free choice be known?” in C. Bicchieri, R. Jeffrey and B. Skyrms (eds.), *The logic of strategy*, Oxford University Press, pp. 163– 174.

Ginet, Carl 1962. “Can the will be caused?”, *The Philosophical Review*, 71:49– 55.

Goldman, Alvin 1970. *A theory of human action*, Princeton University Press.

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

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

Harper, William L. 1988. “Causal decision theory and game theory: a classic argument for equilibrium solutions, a defense of weak equilibria, and a limitation for the normal form representation” in W. L. Harper and B. Skyrms (eds.), *Causation in decision, belief change, and statistics,* II, Kluwer Academic Publishers, pp. 246– 266.

Hausman, Daniel 2012. *Preference, value, choice and welfare*, Cambridge University Press.

Joyce, James M. 2002. “Levi on causal decision theory and the possibility of predicting one’s own actions”, *Philosophical Studies*, 110:69–102.

Kadane, Joseph B. and Seidenfeld, Teddy 1999. “Equilibrium, common knowledge, and optimal sequential decisions” in J. B. Kadane, Mark J. Schervish and T. Seidenfeld (eds.), *Rethinking the Foundations of Statistics*, Cambridge University Press, pp. 27– 46.

Ledwig, Marion 2005. “The no probabilities for acts principle”, *Synthese*, 144:171– 180

Levi, Isaac 1986. *Hard choices*, Cambridge University Press.

Levi, Isaac 1997. *The covenant of reason: rationality and the commitments of thought*, Cambridge University Press.

Lewis, David 1973. *Counterfactuals*, Oxford, Basil Blackwell.

Lewis, David 1981. “Causal decision theory”, *Australasian Journal of Philosophy*, 59:5– 30.

Luce, Duncan R. 1959. *Individual choice behavior: a theoretical analysis*, John Wiley and Sons.

Peterson, Martin 2006. “Indeterminate preferences”, *Philosophical Studies*, 130:297– 320.

Rabinowicz, Wlodek 2000. “Backward induction in games: on an attempt at logical reconstruction” in W. Rabinowicz (ed.), *Value and choice: some common themes in decision theory and moral philosophy*, University of Lund Philosophy Reports, pp. 243– 256.

Rabinowicz, Wlodek 2002. “Does practical deliberation crowd out self-prediction?’, *Erkenntnis* 57:91–122.

Rubinstein, Ariel 2012. *Economic fables*, Open Book Publishers.

Rubinstein, Ariel and Salant, Yuval 2008. “Some thoughts on the principle of revealed preference” in A. Caplin and A. Schotter (eds.), *Handbook of economic methodology,* Oxford University Press, pp. 116– 124.

Shackle, George L.S.1958. *Time in Economics*, North-Holland Publishing Company, Amsterdam.

Shick, Frederic 1979. “Self knowledge, uncertainty and choice”, *British Journal for the Philosophy of Science,* 30:235– 252.

Shin, Hyun Song 1992. “Counterfactuals and a theory of equilibrium in games” in C. Bicchieri and M. L. Dalla Chiara (eds.), *Knowledge, belief and strategic interaction,* Cambridge University Press, pp. 397– 413.

Skyrms, Brian 1982. “Causal decision theory”, *Journal of Philosophy*, 79:695– 711.

Sobel, Jordan H. 1986. “Notes on decision theory: old wine in new bottles”, *Australasian Journal of Philosophy*, 64:407– 437.

Spohn, Wolfgang 1977. “Where Luce and Krantz do really generalize Savage's decision model”, *Erkenntnis*, 11:113– 134.

Spohn, Wolfgang 1999. *Strategic Rationality,* Volume 24 of *Forschungsberichte der DFG-Forschergruppe Logik in der Philosophie,* Konstanz University.

Spohn, Wolfgang 2003. “Dependency equilibria and the causal structure of decision and game situations”, *Homo Oeconomicus*, 20:195– 255.

Spohn, Wolfgang 2007. “Dependency equilibria”, *Philosophy of Science*, 74:775– 789.

Spohn, Wolfgang 2010. “From Nash to dependency equilibria” in G. Bonanno, B. Löwe and W. van der Hoek (eds.), *Logic and the foundations of game and decision theory – LOFT8*, Texts in Logic and Games, Springer, pp. 135– 150.

Spohn, Wolfgang 2012. “Reversing 30 years of discussion: why causal decision theorists should one-box”, *Synthese*, 187:95– 122.

Stalnaker, Robert 1968. “A theory of conditionals” in N. Rescher (ed.), *Studies in logical theory,* Oxford, Blackwell, pp. 98– 112.

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

Weirich, Paul 2008. “Causal decision theory” in *Stanford Encyclopedia of Philosophy*, URL: http://plato.stanford.edu/entries/decision-causal/.

Zambrano, Eduardo 2004. “Counterfactual reasoning and common knowledge of rationality in normal form games”, *Topics in Theoretical Economics*, 4 (1), article 8.

[*] This is the second draft of a chapter in a planned book on the Prisoner's Dilemma, edited by Martin Peterson, to be published by Cambridge University Press.

[1] http://en.wikipedia.org/wiki/Harold_Camping_Rapture_prediction.

[2] For a criticism of the view that preferences are not subject to rational scrutiny see Chapter 10 of Hausman (2012).

[3] Throughout this chapter we view the Prisoner's Dilemma as a strategic-form game with ordinal preferences as follows: $N=\{1,2\}$, $S_1=S_2=\{C,D\}$, $O=\{o_1,o_2,o_3,o_4\}$, $z(C,C)=o_1$, $z(C,D)=o_2$, $z(D,C)=o_3$, $z(D,D)=o_4$, Player 1's ranking of the outcomes is $o_3 \succ_1 o_1 \succ_1 o_4 \succ_1 o_2$ (where $\succ$ denotes strict preference, that is, $x \succ y$ if and only if $x \succsim y$ and not $y \succsim x$) and Player 2's ranking is $o_2 \succ_2 o_1 \succ_2 o_4 \succ_2 o_3$. A preference relation $\succsim$ over the set of outcomes *O* can also be represented by means of an *ordinal utility function* $U : O \to \mathbb{R}$ (where $\mathbb{R}$ denotes the set of real numbers) which satisfies the property that, for any two outcomes o and o', $U(o) \geq U(o')$ if and only if $o \succsim o'$. In Figure 1 we have replaced each outcome with a pair of numbers, where the first is the utility of that outcome for Player 1 and the second is Player 2's utility.

We take preferences over the outcomes as primitives (and utility functions merely as tools for representing those preferences). Thus we are not following the *revealed preference* approach, where observed choices are the primitives and preferences (or utility) are a derived notion:

> "In revealed-preference theory, it isn't true [...] that Pandora chooses *b* rather than *a* because she prefers *b* to *a*. On the contrary, it is because Pandora chooses *b* rather than *a* that we say that Pandora prefers *b* to *a*, and assign *b* a larger utility." (Binmore, 2011, p. 19.)

Thus in the Prisoner's Dilemma game of Figure 1,

> "Writing a larger payoff for Player 1 in the bottom-left cell of the payoff table than in the top-left cell is just another way of registering that Player 1 would choose *D* if she knew that Player 2 were going to choose *C*. [W]e must remember that Player 1 doesn't choose *D* because she then gets a larger payoff. Player 1 assigns a larger payoff to [the outcome associated with] (*D*,*C*) than to [the outcome associated with] (*C*,*C*) because she would choose the former if given the choice." (Binmore, 2011, pp. 27-28, with minor modifications to adapt the quotation to the notation used in Figure 1.)

For a criticism of (various interpretations of) the notion of revealed preference see Chapter 3 of Hausman (2012); see also Rubinstein and Salant (2008).

[4] It is important to note, however, that the payoff functions are taken to be purely ordinal and one could replace π_i with any other function obtained by composing π_i with an arbitrary strictly increasing function on the set of real numbers. In the literature it is customary to impose a stronger assumption on players' preferences, namely that each player has a complete and transitive preference relation on the set of probability distributions over the set of outcomes *O* which satisfies the axioms of Expected Utility. For our purposes this stronger assumption is not needed.

[5] Thus the relation $\mathcal{B}_i$ can also be viewed as a function $\mathcal{B}_i : \Omega \to 2^{\Omega}$; such functions are called *possibility correspondences* in the literature. For further details the reader is referred to Battigalli and Bonanno (1999).

[6] It should be noted, however, that a large literature – that originates in Aumann (1987) – defines rationality in strategic-form games using the models described above, without enriching them with an explicit

framework for counterfactuals. However, as Shin (1992, p. 412) notes "If counterfactuals are not explicitly invoked, it is because the assumptions are buried implicitly in the discussion." We shall return to this point in Section 6.

[7] On the other hand, counterfactuals have been explored extensively in the context of dynamic games. See Bonanno (2013*a*) for a general discussion and relevant references.

[8] For example, the restriction that if $\omega \in E$ then $f(\omega, E) = \{\omega\}$.

[9] Found on the web site http://xkcd.com/1170.

[10] See, for example, Arló-Costa and Bicchieri (2007), Battigalli *et al* (2013), Board (2004), Bonanno (2011), Clausing (2004), Halpern (1999, 2001), Rabinowicz (2000), Stalnaker (1996). For a critical discussion of this approach see Bonanno (2103*a*).

[11] Alternatively, one could remove the initial beliefs $\{\mathcal{B}_i\}_{i \in N}$ from the definition of an extended model and recover them from the function f_i by taking $f_i(\omega, \Omega)$ to be the set of states that player *i* – initially – considers possible at state ω. There are further consistency properties that are usually imposed: (1) if $E \neq \varnothing$ then $f_i(\omega, E) \neq \varnothing$, (2) if $\mathcal{B}_i(\omega) \cap E \neq \varnothing$ then $f_i(\omega, E) = \mathcal{B}_i(\omega) \cap E$ and (3) if $E \subseteq F$ and $f_i(\omega, F) \cap E \neq \varnothing$ then $f_i(\omega, E) = f_i(\omega, F) \cap E$. For a more detailed discussion see Bonanno (2013*a*).

[12] The principles that were introduced by Alchourrón *et al* (1985), which pioneered the vast literature on the so called AGM theory of belief revision.

[13] There are various formulations of causal decision theory: see Gibbard and Harper (1978), Lewis (1981), Skyrms (1982) and Sobel (1986). For an overview see Weirich (2008).

[14] IND_2^{subj} implies that $\bigcup_{\omega'' \in \bigcup_{\omega' \in \mathcal{B}_i(\omega)} f(\omega',[x])} \{\sigma_{-i}(\omega'')\} = \bigcup_{\omega' \in \mathcal{B}_i(\omega)} \{\sigma_{-i}(\omega')\}$ which coincides with IND_1^{subj} if one takes $f_i(\omega,[x]) = \bigcup_{\omega' \in \mathcal{B}_i(\omega)} f(\omega',[x])$.

[15] Although in strategic-form games the two approaches can be considered to be equivalent, this is not so for dynamic games, where the "objective" approach may be too restrictive. This point is discussed in Bonanno (2013*a*).

[16] The models used by Aumann (1987, 1995) make use of knowledge, that is, of necessarily correct beliefs. We refer to these models as epistemic, reserving the term 'doxastic' for models that use the more general notion of belief, which allows for the possibility of error. The models discussed in this chapter are the more general doxastic models.

[17] Recall that $\mathcal{B}_i(\omega)$ is the set of states that player *i* considers possible at state *ω*; recall also the assumption that $\sigma_i(\bullet)$ is constant on $\mathcal{B}_i(\omega)$, that is, for every $\omega' \in \mathcal{B}_i(\omega)$, $\sigma_i(\omega') = \sigma_i(\omega)$.

[18] When payoffs are taken to be von Neumann-Morgenstern payoffs and the beliefs of player *i* at state ω are represented by a probability distribution $\mathbf{p}_{i,\omega} : \Omega \rightarrow [0,1]$ (assuming that Ω is a finite set) whose support coincides with $\mathcal{B}_i(\omega)$ (that is, $\mathbf{p}_{i,\omega}(\omega') > 0$ if and only if $\omega' \in \mathcal{B}_i(\omega)$) then the choice of player *i* at state ω is defined to be rational if and only if it maximizes player *i*'s expected payoff at state ω, that is, if and only if there is no strategy s_i of player *i* such that $\sum_{\omega' \in \mathcal{B}_i(\omega)} \mathbf{p}_{i,\omega}(\omega')\ \pi_i\left(s_i, \sigma_{-i}(\omega')\right) \ > \sum_{\omega' \in \mathcal{B}_i(\omega)} \mathbf{p}_{i,\omega}(\omega')\ \pi_i\left(\sigma_i(\omega), \sigma_{-i}(\omega')\right)$.

[19] Zambrano uses probabilistic beliefs: for every $\omega \in \Omega$, $\mathbf{p}_{i,\omega} : \Omega \rightarrow [0,1]$ is a probability distribution over Ω that represents the beliefs of player *i* at state ω. Our set $\mathcal{B}_i(\omega)$ corresponds to the support of $\mathbf{p}_{i,\omega}$. Zambrano's definition is as follows: player *i* is W-rational at state ω if there is no strategy s_i of player *i* such that $\sum_{\omega' \in \Omega} \mathbf{p}_{i,f_i(\omega,s_i)}(\omega')\ \pi_i\left(s_i, \sigma_j(\omega')\right) > \sum_{\omega' \in \Omega} \mathbf{p}_{i,f_i(\omega,s_i)}(\omega')\ \pi_i\left(\sigma_i(\omega), \sigma_j(\omega')\right)$.

[20] Given that Zambrano postulates probabilistic beliefs, he expresses this condition as follows: $\mathrm{marg}_{S_j}\, p_{i,\omega}(\bullet) = \mathrm{marg}_{S_j}\, p_{i,f_i(\omega,s_i)}(\bullet)$.

[21] Recall the assumption that a player always knows her chosen strategy, that is, for every $\omega' \in \mathcal{B}_i(\omega)$, $\sigma_i(\omega') = \sigma_i(\omega)$ and thus $-$ since we are considering a strategy $s_i \neq \sigma_i(\omega)$ $-$ it must be the case that $f(\omega', s_i) \neq \omega'$.

[22] This set can also be written as $\bigcup_{\omega'' \in \bigcup_{\omega' \in \mathcal{B}_i(\omega)} f(\omega', s_i)} \left\{\sigma_{-i}(\omega'')\right\}$.

[23] Like Zambrano, Board assumes that payoffs are von Neumann-Morgenstern payoffs and beliefs are probabilistic: for every $\omega \in \Omega$, $\mathbf{p}_{i,\omega}$ is a probability distribution with support $\mathcal{B}_i(\omega)$ that represents the probabilistic beliefs of player *i* at state ω. Board defines player *i* to be causally rational at state ω if there is no strategy s_i that would yield a higher expected payoff if chosen instead of $\sigma_i(\omega)$, that is, if there is no $s_i \in S_i$ such that $\sum_{\omega' \in \mathcal{B}_i(\omega)} \mathbf{p}_{i,\omega}(\omega')\ \pi_i\left(s_i, \sigma_{-i}\left(f(\omega', s_i)\right)\right) > \sum_{\omega' \in \mathcal{B}_i(\omega)} \mathbf{p}_{i,\omega}(\omega')\ \pi_i\left(\sigma_i(\omega), \sigma_{-i}(\omega')\right)$. There is no clear qualitative counterpart to this definition, because of the lack of any constraints that relate $\bigcup_{\omega' \in \mathcal{B}_i(\omega)} \left\{\sigma_{-i}(f(\omega', s_i))\right\}$ to $\bigcup_{\omega' \in \mathcal{B}_i(\omega)} \left\{\sigma_{-i}(\omega')\right\}$. Board (2006, p. 16) makes this point as follows: "since each state describes what each player does as well as what her opponents do, the player will change the state if she changes her choice. There is no guarantee that her opponents will do the same in the new state as they did in the original state."

[24] Board presents this as an objective condition on the selection function (if $\omega' = f(\omega, s_i)$ then $\sigma_{-i}(\omega') = \sigma_{-i}(\omega)$) assumed to hold at every state (and thus imposed as an axiom), but then acknowledges (p. 12) that "it is players' beliefs in causal independence rather than causal independence itself that drives the result."

[25] Similar observations can be found in Schick (1979), Gilboa (1999), Kadane and Seidenfeld (1999); for a discussion and further references see Ledwig (2005). It should be noted, however, that this view has been criticized by several authors: see, for example, Joyce (2002), Rabinowicz (2002) and Peterson (2006); Luce (1959) also claimed that it sometimes makes sense to assign probabilities to one’s own choices.