Game theory with translucent players

Joseph Y. Halpern · Rafael Pass

Accepted: 7 May 2018 / Published online: 18 June 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract  A traditional assumption in game theory is that players are opaque to one another—if a player changes strategies, then this change in strategies does not affect the choice of other players' strategies. In many situations this is an unrealistic assumption. We develop a framework for reasoning about games where the players may be translucent to one another; in particular, a player may believe that if she were to change strategies, then the other player would also change strategies. Translucent players may achieve significantly more efficient outcomes than opaque ones. Our main result is a characterization of strategies consistent with appropriate analogues of common belief of rationality. Common Counterfactual Belief of Rationality (CCBR) holds if (1) everyone is rational, (2) everyone counterfactually believes that everyone else would still be rational even if she were to switch strategies, (3) everyone counterfactually believes that everyone else is rational (i.e., all players $i$ believe that everyone else would still be rational even if $i$ were to switch strategies), (4) everyone counterfactually believes that everyone else is ratio-

A short preliminary version of this paper appeared in the Fourteenth Conference on Theoretical Aspects of Rationality and Knowledge (TARK), 2013, pp. 216–221. We thank the anonymous referees and the associate editor for the useful comments.

Halpern is supported in part by NSF grants IIS-0812045, IIS-0911036, and CCF-1214844, by AFOSR grant FA9550-08-1-0266, and by ARO grant W911NF-09-1-0281. Pass is supported in part by a Alfred P. Sloan Fellowship, Microsoft New Faculty Fellowship, NSF Award CNS-1217821, NSF CAREER Award CCF-0746990, NSF Award CCF-1214844, AFOSR YIP Award FA9550-10-1-0093, and DARPA and AFRL under contract FA8750-11-2-0211. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the US Government.

Joseph Y. Halpern  halpern@cs.cornell.edu
Rafael Pass  rafael@cs.cornell.edu

Department of Computer Science, Cornell University, Ithaca, NY 14853, USA
nal, and counterfactually believes that everyone else is rational, and so on. CCBR characterizes the set of strategies surviving iterated removal of minimax-dominated strategies, where a strategy $\sigma$ for player $i$ is minimax dominated by $\sigma'$ if the worst-case payoff for $i$ using $\sigma'$ is better than the best possible payoff using $\sigma$.

1 Introduction

Two large firms 1 and 2 need to decide whether to cooperate ($C$) or sue ($S$) the other firm. Suing the other firm always has a small positive reward, but being sued induces a high penalty $p$; more precisely, $u(C, C) = (0, 0); u(C, S) = (-p, r); u(S, C) = (r - p, r - p)$. In other words, we are considering an instance of Prisoner’s Dilemma.

The firms agree to cooperate. But there is a catch. Before acting, each firm needs to discuss their decision with its board. Although these discussions are held behind closed doors, there is always the possibility of a change in plans being “leaked”; as a consequence, if one company decides to change its course of action, the other company may do so as well. (If there is no change in plan, there is no leak.) Furthermore, both companies are aware of this fact. In other words, the players are translucent to one another.

In such a scenario, it may well be rational for both companies to cooperate. For instance, consider the following situation.

- Firm $i$ believes that a change in plan is leaked to firm $3 - i$ with probability $\epsilon$.
- Firm $i$ believes that if the other firm $3 - i$ finds out that $i$ is defecting, then $3 - i$ will also defect. That is, firm $i$ believes that firm $3 - i$’s mixed strategy is a function of its own strategy.
- Finally, $p \epsilon > r$ (i.e., the penalty for being sued is significantly higher than the reward of suing the other company).

Neither firm defects, since defection is noticed by the other firm with probability $\epsilon$, which (according to their beliefs) leads to a harsh punishment. Thus, the possibility of a changed in plan being leaked to the other player allows the players to significantly improve social welfare in equilibrium. (This suggests that it may be mutually beneficial for two countries to spy on each other!)

Even if the Prisoner’s Dilemma is not played by corporations but by individuals, each player may believe that if he chooses to defect, his “guilt” over defecting may be visible to the other player. [Indeed, facial and body cues such as increased pupil size are often associated with deception; see e.g., (Ekman and Friesen 1969)]. There is indeed evidence that players are quite good at reading and responding to cues in Prisoner’s Dilemma-like games: the frequency with which players make the same choice (both cooperating or both defecting) is significantly higher than the frequency with which they make different choices (Kalay 2003).

Thus, again, the players may choose to cooperate out of fear that if they defect, the other player may detect it and act on it. (Furthermore, even if player 2 actually cannot/does not react to a switch in player 1’s strategy, player 1 may believe that this switch is visible and that player 2 may react to it.)
A similar phenomenon arises in evolutionary biology. It has been observed that in biological settings, the individuals that a particular individual \( i \) encounters are correlated with the strategy \( \sigma \) that individual \( i \) uses and, in particular, the fraction of these individuals that use some particular strategy \( \sigma' \) is a function of \( \sigma \). Essentially, this can be viewed as saying that the mixed strategy used by \( i \)’s opponent is a function of \( i \)’s strategy. Eshel and Cavalli-Sforza (1982) define a notion of evolutionarily stable strategy that takes this into account, and show that it can be used to explain the evolution on cooperation in Prisoner’s Dilemma-like settings.\(^1\)

Our goal is to model a large class of games where a player \( i \) may believe that, if she were to switch strategies, the other players might switch strategies as well, and to consider appropriate solution concepts for such games. We take a Bayesian approach: Each player has a (subjective) probability distribution (describing the player’s beliefs) over the states of the world and wants to maximize expected utility.\(^2\) Traditionally, a player \( i \) is said to be rational in a state \( \omega \) if the strategy \( \sigma_i \) that \( i \) plays at \( \omega \) is a best response to the strategy profile \( \mu_{-i} \) of the other players induced by \( i \)’s beliefs in \( \omega \);\(^3\) that is, \( u_i(\sigma_i, \mu_{-i}) \geq u_i(\sigma'_i, \mu_{-i}) \) for all alternative strategies \( \sigma'_i \) for \( i \). In our setting, things are more subtle, since player \( i \)’s beliefs about what the other players are doing may change if she switches strategies. The notion of “best response” must take this into account.

But how exactly would player \( i \)’s beliefs change if she were to switch strategies? We started this paper with one specific model of belief change in response to a possible leakage of information in Prisoner’s Dilemma. Solan and Yariv (2004) considered another specific model of leakage in their games with espionage. A game with espionage is a two-player extensive-form game that extends an underlying normal-form game by adding a step where player 1 can purchase some noisy information about player 2’s planned move.\(^4\) If we move beyond two-player games, the situation gets even more complicated, since there may be a chain reaction of players reacting to changes in strategies, with players reacting to each change in strategy. Moreover, changes in beliefs may not be due to factors other than information leakage. For example, the analysis of biological settings by Eshel and Cavalli-Sforza (1982) provides quite a different model of belief change, due to correlations between players’ actions.

Rather than considering a class of extensive-form games that explicitly describes one particular process of belief change (as, for example, Solan and Yariv (2004) did),

---

1. We thank Sergiu Hart for pointing us to this paper.
2. The notion of translucency makes perfect sense even if we consider decision rules other than expected utility maximization. We could easily modify the technical details presented in Sect. 2 to handle other decision rules. We focus on expected utility maximization so as to be able to relate our results to traditional solution concepts such as rationalizability.
3. Formally, we assume that \( i \) has a distribution on states, and at each state, a pure strategy profile is played; the distribution on states clearly induces a distribution on strategy profiles for the players other than \( i \), which we denote \( \mu_{-i} \). More generally, if \( X = X_1 \times \cdots \times X_n \) and \( x \) \( \in \) \( X \), we use \( x_{-i} \) to denote \( (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \), and \( X_{-i} \) to denote \( X_1 \times \cdots \times X_{i-1} \times X_{i+1} \cdots \times X_n \).
4. Of course, adding a cost to leakage introduces new forms of strategic behavior. In our setting, the information is free—we deliberately do not consider the strategic implications of obtaining information—and all players may be translucent, so the situation is more symmetric.
we stick with the simple setting of normal-form games, but augment a model of the
game with an explicit description of how players’ beliefs change, using counterfactuals
(Lewis 1973; Stalnaker 1968). Formally, associated with each state of the world \( \omega \),
each player \( i \), and strategy \( \sigma_i \), there is a “closest-state” \( f(\omega, i, \sigma'_i) \) where player \( i \)
plays \( \sigma'_i \). Note that if \( i \) changes strategies, then this change may start a chain reaction
among the other players. We can think of \( f(\omega, i, \sigma'_i) \) as the steady-state outcome of
this process: the state that would result if \( i \) switched strategies to \( \sigma'_i \). Let \( \mu_f(\omega, i, \sigma'_i) \)
be the distribution on strategy profiles of \(-i\) (the players other than \( i \)) induced by \( i\)'s
beliefs at \( \omega \) about the steady-state outcome of this process. We say that \( i \) is rational at a
state \( \omega \) where \( i \) plays \( \sigma_i \) and has beliefs \( \mu_i \) if \( u_i(\sigma_i, \mu_{-i}) \geq u_i(\sigma'_i, \mu_f(\omega, i, \sigma'_i)) \) for every
alternative strategy \( \sigma'_i \) for \( i \). Note that we have required the closest-state function to
be deterministic, returning a unique state, rather than a distribution over states. While
this may seem incompatible with the motivating scenario, in fact, it is not. By taking
a rich enough representation of states, we can assume (essentially, without loss of
generality in our setting) that a state contains enough information about players to
resolve uncertainty about what strategies they would use if one player were to switch
strategies.

To see why this should be the case, in our motivating example, suppose that each
firm believes that, if it decides to defect, the decision will leak with probability .3. We
can model this by assuming that, corresponding to each action profile, there are
four possible states, depending on whether each firm’s decision leaks. In each of these
state, there is an obvious (unique) closest state. For example, in the state \( \omega \) where
both players cooperate, firm 1’s decision leaks, and firm 2’s decision does not (which
has probability .21), the unique state closest to \( \omega \) where firm 1 defects is the state
where both firms defect (and all other features are identical), since firm 1’s defection
leaks and results in firm 2 defecting as well, while the unique state closest to \( \omega \) where
firm 2 defects is the state where firm 2 defects but player 1 does not (since firm 2’s
decision does not leak in state \( \omega \)). More generally, given a model where the closest-
state function returns a probability distribution over states, we can convert it to a model
where the closest-state function returns a unique state that is equivalent to the original
model in the sense that the same statements in the language of interest are true in both
models. See Section 2.3 for more discussion.

While allowing arbitrary closest-state functions means that we are considering a
rather large class of models, we note that this approach is in the spirit of allowing
agents arbitrary beliefs about other agents’ strategies and arbitrary utility functions in
a game. Of course, it would be of interest to consider specific closest-state functions,
corresponding to various belief-change processes. It would also be of interest to limit
the closest-state function by imposing rationality considerations. This, in fact, is what
we do when considering solution concepts.

We consider two solution concepts in a setting with translucent players, and provide
epistemic characterizations of them. The first is an analogue of rationalizability (Bern-
heim 1984; Pearce 1984). To give a sense of what we do, we need some definitions.

---

5 We are implicitly assuming that \( i \)'s strategy \( \sigma'_i \) is being held fixed during this chain reaction, so that in
the state \( f(\omega, i, \sigma'_i) \), \( i \) is playing strategy \( \sigma'_i \).
We say that a player \( i \) counterfactually believes \( \varphi \) at \( \omega \) if \( i \) believes \( \varphi \) holds even if \( i \) were to switch strategies. Common Counterfactual Belief of Rationality (CCBR) holds if (1) everyone is rational, (2) everyone counterfactually believes that everyone else is rational (i.e., all players \( i \) believe that everyone else would still be still rational even if \( i \) were to switch strategies), (3) everyone counterfactually believes that everyone else is rational, and counterfactually believes that everyone else is rational, and so on.

Our main result is a characterization of strategies consistent with CCBR. Roughly speaking, these results can be summarized as follows:

- If the closest-state function respects “unilateral deviations”—when \( i \) switches strategies, the strategies and beliefs of players other than \( i \) remain the same—then CCBR characterizes the set of rationalizable strategies.
- If the closest-state function can be arbitrary, CCBR characterizes the set of strategies that survive iterated removal of minimax-dominated strategies: a strategy \( \sigma_i \) is minimax dominated for \( i \) if there exists a strategy \( \sigma'_i \) for \( i \) such that \( \min_{\mu'_{-i}} u_i(\sigma'_i, \mu'_{-i}) > \max_{\mu_{-i}} u_i(\sigma_i, \mu_{-i}) \); that is, \( u_i(\sigma'_i, \mu'_{-i}) > u_i(\sigma_i, \mu_{-i}) \) no matter what the strategy profiles \( \mu_{-i} \) and \( \mu'_{-i} \) are.

We then consider analogues of Nash equilibrium in our setting, and show that individually rational strategy profiles that survive iterated removal of minimax-dominated strategies characterize such equilibria.

Note that in our approach, each player \( i \) has a belief about how the other players’ strategies would change if \( i \) were to change strategies, but we do not require \( i \) to explicitly specify how he would respond to other people changing strategies. The latter approach, of having each player specify how she responds to her opponents’ actions, goes back to von Neumann and Morgenstern (1947, pp. 105–106):

Indeed, the rules of the game \( \Gamma \) prescribe that each player must make his choice (his personal move) in ignorance of the outcome of the choice of his adversary. It is nevertheless conceivable that one of the players, say 2, “finds out”; i.e., has somehow acquired the knowledge as to what his adversary’s strategy is. The basis for this knowledge does not concern us; it may (but need not) be experience from previous plays.

Von Neumann and Morgenstern’s analysis suggests that one should do a single round of removal of minimax-dominated strategies. This approach was further explored and formalized by by Howard (1971) in the 1960s. In Howard’s approach, players pick a “meta-strategy” that takes as input the strategy of other players. It led to complex formalisms involving infinite hierarchies of meta-strategies: at the lowest level, each player specifies a strategy in the original game; at level \( k \), each player specifies a “response rule” (i.e., a meta-strategy) to other players’ \((k-1)\)-level response rules. Such hierarchical structures have not proven useful when dealing with applications. Since we do not require players to specify reaction rules, we avoid the complexities of this approach.

Program equilibria (Tennenholz 2004) and conditional commitments (Kalai et al. 2010) provide a different approach to avoiding infinite hierarchies. Roughly speaking, each player \( i \) simply specifies a program \( \Pi_i \); player \( i \)’s action is determined by running \( i \)’s program on input the (description of) the programs of the other players; that is, \( i' \)
action is given by $\Pi_i(\Pi_{-i})$. Tennenholz (2004) and Kalai et al. (2010) show that every (correlated) individually rational outcome can be sustained in a program equilibrium. Their model, however, assumes that player’s programs (which should be interpreted as their “plan of action”) are commonly known to all players. We dispense with this assumption. It is also not clear how to define common belief of rationality in their model; the study of program equilibria and conditional commitments has considered only analogues of Nash equilibrium.

Perhaps most closely related to our model is a paper by Spohn (2003) that studies a generalization of Nash equilibrium called dependency equilibrium, where players’ conjectures are described as “conditional probabilities”: for each action $a_1$ of player 1, player 1 may have a different belief about the action of player 2. Spohn’s notion of dependency equilibrium is essentially equivalent to the analogues of Nash equilibrium in our setting characterized in Sect. 4 (see Theorem 4.4). Tillio et al. (2014) also allow for the possibility that agent $i$’s beliefs regarding the actions of other players may change as $i$ conditions on different actions (their type functions play the same role as our closest-state functions), and observe that it leads to a notion of rationality that allows dominated actions to be rational (although they do not consider solution concepts). Independently of our work, Salcedo (2013), defined a notion of counterfactual rationalizability that replaces beliefs (over actions) by conjectures described as conditional probabilities, as in Spohn (2003). Salcedo also defines a notion of minimax domination (which he calls absolute domination), and characterizes counterfactual rationalizability in terms of strategies surviving iterated deletion of minimax-dominated strategies (although he does not give an epistemic characterization). Thus, a strategy is counterfactually rationalizable in Salcedo’s sense if and only if it is consistent with CCBR. Capraro (2013) also independently introduced minimax-dominated strategies (calling them super-dominated strategies), and considered iterated deletion of minimax-dominated strategies, although he did not use them to provide a characterization of a solution concept.

Counterfactuals have been explored in a game-theoretic setting; see, for example, (Aumann 1995; Halpern 1999; Samet 1996; Stalnaker 1996; Zambrano 2004). However, all these papers considered only structures where, in the closest state where $i$ changes strategies, all other players’ strategies remain the same; thus, these approaches are not applicable in our context.

We can view our framework as an attempt to formalize quasi-magical thinking (Shafir and Tversky 1992), the kind of reasoning that is supposed to motivate those people who believe that the others’ reasoning is somehow influenced by their own thinking, even though they know that there is no causal relation between the two. Quasi-magical thinking has also been formalized by Masel (2007) in the context of the Public Goods game and by Daley and Sadowski (2014) in the context of symmetric $2 \times 2$ games.

### 2 Counterfactual structures

Given a game $\Gamma$, let $\Sigma_i(\Gamma)$ denote player $i$’s pure strategies in $\Gamma$ (we occasionally omit the parenthetical $\Gamma$ if it is clear from context or irrelevant). Since our focus here is on
normal-form games, a strategy is simply an action. However, our definitions extend naturally to extensive-form games.

To reason about the game $\Gamma$, we consider a class of Kripke structures corresponding to $\Gamma$. For simplicity, we here focus on finite structures. A finite probability structure $M$ appropriate for $\Gamma$ is a tuple $(\Omega, s, PR_1, \ldots, PR_n)$, where $\Omega$ is a finite set of states; $s$ associates with each state $\omega \in \Omega$ a pure strategy profile $s(\omega)$ in the game $\Gamma$; and, for each player $i$, $PR_i$ is a probability assignment that associates with each state $\omega \in \Omega$ a probability distribution $PR_i(\omega)$ on $\Omega$, such that

1. $PR_i(\omega)([[s_i(\omega)]_M) = 1$, where for each strategy $\sigma_i$ for player $i$, $[[\sigma_i]]_M = \{\omega : s_i(\omega) = \sigma_i\}$, where $s_i(\omega)$ denotes player $i$’s strategy in the strategy profile $s(\omega)$;
2. $PR_i(\omega)([[PR_i(\omega), i]]_M) = 1$, where for each probability measure $\pi$ and player $i$, $[[\pi, i]]_M = \{\omega : PR_i(\omega) = \pi\}$.

These assumptions say that player $i$ assigns probability 1 to his actual strategy and beliefs.

To deal with counterfactuals, we augment probability structures with a “closest-state” function $f$ that associates with each state $\omega$, player $i$, and strategy $\sigma_i$, a state $f(\omega, i, \sigma_i)$ where player $i$ plays $\sigma_i$; if $\sigma_i$ is already played in $\omega$, then the closest state to $\omega$ where $\sigma_i$ is played is $\omega$ itself. Formally, a finite counterfactual structure $M$ appropriate for $\Gamma$ is a tuple $(\Omega, s, f, PR_1, \ldots, PR_n)$, where $(\Omega, s, PR_1, \ldots, PR_n)$ is a probability structure appropriate for $\Gamma$ and $f$ is a “closest-state” function. We require that if $f(\omega, i, \sigma_i) = \omega'$, then

1. $s_i(\omega') = \sigma_i'$;
2. if $\sigma_i' = s_i(\omega)$, then $\omega' = \omega$.

Given a probability assignment $PR_i$ for player $i$, we define $i$’s counterfactual belief at state $\omega$ (“what $i$ believes would happen if he switched to $\sigma_i'$ at $\omega$) as

$$PR_{i, \sigma_i}(\omega') = \sum_{\omega'' \in \Omega : f(\omega'', i, \sigma_i')} PR_i(\omega)(\omega'').$$

Thus, if player $i$ were to play $\sigma_i'$, the probability that $i$’s would assign at state $\omega$ to the state $\omega'$ is just the probability (according to the probability measure $Pr_i(\omega)$) of the event “the closest world where player $i$ plays $\sigma_i'$ is $\omega'$” (this event consists of the state $\omega''$ such that $f(\omega'', i, \sigma_i') = \omega'$). Note that the conditions above imply that each player $i$ knows what strategy he would play if he were to switch; that is, $PR_{i, \sigma_i}(\omega)([[\sigma_i]]_M = 1$.

Let $Supp(\pi)$ denote the support of the probability measure $\pi$. Note that $Supp(PR_{i, \sigma_i})(\omega) = \{f(\omega', i, \sigma_i') : \omega' \in Supp(PR_i(\omega))\}$. Moreover, it is almost

---

6 All the results in the paper go through without change if we replace “finite” by “countable” everywhere. While we believe that our basic results apply to arbitrary structures, extending them to arbitrary structures requires dealing with measurability issues, which would distract us from the main points that we are trying to make here. On the other hand, restricting to finite or countable structures means that we cannot model situations where a player $i$ has no idea of what $j$’s probability distribution on other players’ strategies is, and thus wants to consider all probability distributions possible.
immediate from the definition that if \( PR_i(\omega) = PR_i(\omega') \), then \( PR_{i,\sigma_i}(\omega) = PR_{i,\sigma_i}(\omega') \) for all strategies \( \sigma_i' \) for player \( i \), so \( PR_i(\omega)\llbracket PR_{i,\sigma_i}(\omega), i \rrbracket_M = 1 \). But it does not in general follow that \( i \) knows what his counterfactual beliefs at \( \omega \) would be if he were to switch strategies, that is, it may not be the case that for all strategies \( \sigma_i' \) for player \( i \), \( PR_{i,\sigma_i}(\omega)\llbracket PR_{i,\sigma_i}(\omega), i \rrbracket_M = 1 \). Suppose that we think of a state as representing each player’s ex ante view of the game. The fact that \( s_i(\omega) = \sigma_i \) should then be interpreted as “\( i \) intends to play \( \sigma_i \) at state \( \omega \)” With this view, suppose that \( \omega \) is a state where \( s_i(\omega) \) is a conservative strategy, while \( \sigma_i' \) is a rather reckless strategy. It seems reasonable to expect that \( i \)’s subjective beliefs regarding the likelihood of various outcomes may depend in part on whether he is in a conservative or reckless frame of mind. We can think of \( PR_{i,\sigma_i}(\omega)(\omega') \) as the probability that \( i \) ascribes, at state \( \omega \), to \( \omega' \) being the outcome of \( i \) switching to strategy \( \sigma_i' \); thus, \( PR_{i,\sigma_i}(\omega)(\omega') \) represents \( i \)’s evaluation of the likelihood of \( \omega' \) when he is in a conservative frame of mind. This may not be the evaluation that \( i \) uses in states in the support \( PR_{i,\sigma_i}(\omega) \); at all these states, \( i \) is in a “reckless” frame of mind. Moreover, there may not be a unique reckless frame of mind, so that \( i \) may not have the same beliefs at all the states in the support of \( PR_{i,\sigma_i}(\omega) \).

\( M \) is a strongly appropriate counterfactual structure if it is an appropriate counterfactual structure and, at every state \( \omega \), every player \( i \) knows his counterfactual beliefs (i.e., \( PR_{i,\sigma_i}(\omega)\llbracket PR_{i,\sigma_i}(\omega), i \rrbracket_M = 1 \) for all strategies \( \sigma_i' \)). As the example above suggests, strong appropriateness is a nontrivial requirement. As we shall see, however, our characterization results hold in both appropriate and strongly appropriate counterfactual structures.

Note that even in strongly appropriate counterfactual structures, we may not have \( PR_i(f(\omega, i, \sigma_i')) = PR_{i,\sigma_i}(\omega) \). We do have \( PR_i(f(\omega, i, \sigma_i')) = PR_{i,\sigma_i}(\omega) \) in strongly appropriate counterfactual structures if \( f(\omega, i, \sigma_i') \) is in the support of \( PR_{i,\sigma_i}(\omega) \) (which will certainly be the case if \( \omega \) is in the support of \( PR_i(\omega) \)). To see why we may not want to have \( PR_i(f(\omega, i, \sigma_i')) = PR_{i,\sigma_i}(\omega) \) in general, even in strongly appropriate counterfactual structures, consider the example above again. Suppose that, in state \( \omega \), although \( i \) does not realize it, he has been given a drug that affects how he evaluates the state. He thus ascribes probability 0 to \( \omega \). In \( f(\omega, i, \sigma_i') \) he has also been given the drug, and the drug in particular affects how he evaluates outcomes. Thus, \( i \)’s beliefs in the state \( f(\omega, i, \sigma_i') \) are quite different from his beliefs in all states in the support of \( PR_{i,\sigma_i}(\omega) \).

### 2.1 Logics for counterfactual games

Let \( L(\Gamma) \) be the language where we start with \textit{true} and the primitive proposition \( RAT_i \) and \( \text{play}_i(\sigma_i) \) for \( \sigma_i \in \Sigma_i(\Gamma) \), and close off under the modal operators \( B_i \) (player \( i \) believes) and \( B_i^* \) (player \( i \) counterfactually believes) for \( i = 1, \ldots, n \). \( CB \) (common belief), and \( CB^* \) (common counterfactual belief), conjunction, and negation. We think of \( B_i \varphi \) as saying that “\( i \) believes \( \varphi \) holds with probability 1” and \( B_i^* \varphi \) as saying

\( \odot \) Springer
“i believes that \( \varphi \) holds with probability 1, even if i were to switch strategies”. Let \( \mathcal{L}^0(\Gamma) \) be defined exactly like \( \mathcal{L}(\Gamma) \) except that we exclude the “counterfactual” modal operators \( B^n \) and \( CB^n \). We often omit the parenthetical \( \Gamma \), simply writing \( \mathcal{L} \) and \( \mathcal{L}^0 \), whenever \( \Gamma \) is clear from the context.

We first define semantics for \( \mathcal{L}_0 \) using probability structures (without counterfactuals). We define the notion of a formula \( \phi \) being true at a state \( \omega \) in a probability structure \( M \) (written \( (M, w) \models \phi \)) in the standard way, by induction on the structure of \( \phi \), as follows:

- \( (M, \omega) \models \text{true} \) (so true is vacuously true).
- \( (M, \omega) \models \text{play}_i(\sigma_i) \iff \sigma_i = s_i(\omega) \).
- \( (M, \omega) \models \neg \phi \iff (M, \omega) \not\models \phi \).
- \( (M, \omega) \models \phi \land \phi' \iff (M, \omega) \models \phi \) and \( (M, \omega) \models \phi' \).
- \( (M, \omega) \models B_i \phi \iff \mathcal{P}_i(\omega)((\mathcal{P}_i \phi)_M) = 1 \), where \( \mathcal{P}_i \phi = \{ \omega : (M, \omega) \models \phi \} \).
- \( (M, \omega) \models \text{RAT}_i \) iff \( s_i(\omega) \) is a best response given player i’s beliefs regarding the strategies of other players induced by \( \mathcal{P}_i \); that is, for every strategy \( \sigma'_i \) for player i,

\[
\sum_{\omega' \in \Omega} \mathcal{P}_i(\omega)(\omega')s_i(\omega), s_{-i}(\omega') \geq \sum_{\omega' \in \Omega} \mathcal{P}_i(\omega)(\omega')u_i(\sigma'_i, s_{-i}(\omega')). \tag{1}
\]

- Let \( EB \varphi \) (“everyone believes \( \varphi \)”) be an abbreviation of \( B_1 \varphi \land \ldots \land B_n \varphi \); and define \( EB^k \varphi \) for all \( k \) inductively, by taking \( EB^1 \varphi \) to be \( EB \varphi \) and \( EB^{k+1} \varphi \) to be \( EB(EB^k \varphi) \).
- \( (M, \omega) \models CB \varphi \iff (M, \omega) \models EB^k \varphi \) for all \( k \geq 1 \).

Semantics for \( \mathcal{L}^0 \) in counterfactual structures is defined in the same way, except that we redefine \( \text{RAT}_i \) to take into account the fact that player i’s beliefs about the strategies of players –i may change if i changes strategies.

- \( (M, \omega) \models \text{RAT}_i \) iff for every strategy \( \sigma'_i \) for player i,

\[
\sum_{\omega' \in \Omega} \mathcal{P}_i(\omega)(\omega')u_i(\sigma'_i, s_{-i}(\omega')) \geq \sum_{\omega' \in \Omega} \mathcal{P}_i(\omega)(\omega')u_i(\sigma'_i, s_{-i}(\omega')). \tag{2}
\]

The inequality (2) is equivalent to

\[
\sum_{\omega' \in \Omega} \mathcal{P}_i(\omega)(\omega')u_i(\sigma'_i, s_{-i}(\omega')) \geq \sum_{\omega' \in \Omega} \mathcal{P}_i(\omega)(\omega')u_i(f(\omega', i, \sigma'_i))).
\]

Note that, in general, (2) is different from requiring that \( s_i(\omega) \) is a best response given player i’s beliefs regarding the strategies of other players induced by \( \mathcal{P}_i \), which is given by (1).

Note that in models without counterfactuals, we continue to define \( \text{RAT}_i \) using (1), as opposed to (2). We use the same symbol for both notions to emphasize that both capture the “natural” notion of rationality in the corresponding settings. For each of our results, we will make clear whether we are working with probability structures.
[in which case $RAT_i$ is defined using (1)] or counterfactual structures [in which case $RAT_i$ is defined using (2)].

To give the semantics for $L$ in counterfactual structures, we now also need to define the semantics of $B_i^*$ and $CB^*$:

- $(M, ω) \models B_i^*ϕ$ iff for all strategies $σ_i' \in Σ_i(Γ)$, $P^c_{i,σ_i}(ω)([ϕ]_M) = 1$.
- $(M, ω) \models CB^*ϕ$ iff $(M, ω) \models (EB^*)^kϕ$ for all $k ≥ 1$.

It is easy to see that, like $B_i$, $B_i^*$ depends only on $i$’s beliefs; as we observed above, if $P^c_{i}(ω) = P^c_{i}(ω')$, then $P^c_{i,σ_i}(ω) = P^c_{i,σ_i}(ω')$ for all $σ_i'$, so $(M, ω) \models B_i^*ϕ$ iff $(M, ω') \models B_i^*ϕ$. It immediately follows that $B_i^*ϕ \Rightarrow B_iB_i^*ϕ$ is valid (i.e., true at all states in all structures).

The following abbreviations will be useful in the sequel. Let $RAT$ be an abbreviation for $RAT_1 \land \ldots \land RAT_n$, and let $play(\vec{σ})$ be an abbreviation for $play_1(σ_1) \land \ldots \land play_n(σ_n)$.

### 2.2 Common counterfactual belief of rationality

We are interested in analyzing strategies being played at states where (1) everyone is rational, (2) everyone counterfactually believes that everyone else is rational (i.e., for every player $i$, $i$ believes that everyone else would still be rational even if $i$ were to switch strategies), (3) everyone counterfactually believes that everyone else is rational, and counterfactually believes that everyone else is rational, and so on. For each player $i$, define the formulas $SRAT^k_i$ (player $i$ is strongly $k$-level rational) inductively, by taking $SRAT^0_i$ to be true and $SRAT^{k+1}_i$ to be an abbreviation of $RAT_i \land B_i^*(\land_j \neq i SRAT^k_j)$.

Let $SRAT^k$ be an abbreviation of $\land^n_{j=1}SRAT^k_j$.

Define $CCBR$ (common counterfactual belief of rationality) as follows:

- $(M, ω) \models CCBR$ iff $(M, ω) \models SRAT^kϕ$ for all $k ≥ 1$.

Note that it is critical in the definition of $SRAT^k_i$ that we require only that player $i$ counterfactually believes that everyone else (i.e., the players other than $i$) are rational, and believe that everyone else is rational, and so on. Player $i$ has no reason to believe that his own strategy would be rational if he were to switch strategies; indeed, $B_i^*(RAT_i)$ can hold only if every strategy for player $i$ is rational with respect to $i$’s beliefs. This is why we do not define $CCBR$ as $CB^*(RAT)$. 7

We also consider the consequence of just common belief of rationality in our setting. Define $WRAT^k_i$ (player $i$ is weakly $k$-level rational) just as $SRAT^k_i$, except that $B_i^*$ is replaced by $B_i$. An easy induction on $k$ shows that $WRAT^{k+1}_i$ implies $WRAT^k_i$ and

---

7 Interestingly, Samet (1996) essentially considers an analogue of $CB^*(RAT)$. This definition does not cause problems in his setting since he considers only events in the past, not events in the future.
that \( \text{WRAT}^k \) implies \( B_i(\text{WRAT}^k) \).\(^8\) It follows that we could have equivalently defined \( \text{WRAT}^k_{i+1} \) as

\[
\text{RAT}_i \land B_i(\land_{j=1}^n \text{WRAT}_j^k).
\]

Thus, \( \text{WRAT}^k_{i+1} \) is equivalent to \( \text{RAT} \land \text{EB}(\text{WRAT}^k) \). As a consequence we have the following:

**Proposition 2.1** \((M, \omega) \models \text{CB}(\text{RAT}) \iff (M, \omega) \models \text{WRAT}^k \) for all \( k \geq 0 \).

### 2.3 Probabilistic vs. deterministic closest-state functions

In this section, we make precise our earlier claim that the assumption that the closest-state function returns a unique state is essentially without loss of generality.\(^9\) For the purposes of this section only, assume that in a counterfactual structure \( M = (\Omega, s, f, \mathcal{PR}_1, \ldots, \mathcal{PR}_n) \), \( f(\omega, i, \sigma'_i) \) is an element of \( \Delta(\Omega) \), the set of probability measures on \( \Omega \). We now assume that \( s(\omega') = \sigma'_i \) for all \( \omega' \in \text{Supp}(f(\omega, i, \sigma'_i)) \) and that \( f(\omega, i, \sigma'_i)(\omega) = 1 \) if \( s(\omega) = \sigma'_i \). We then modify \( \mathcal{PR}_{i,\sigma'_i} \) in the obvious way:

\[
\mathcal{PR}_{i,\sigma'_i}(\omega)(\omega') = \sum_{\omega'' \in \Omega} f(\omega'', i, \sigma'_i)(\omega') \mathcal{PR}_i(\omega)(\omega'').
\]

We call a closest-state function that returns a unique state a *deterministic* closest-state function, and call the generalization considered in this section a *probabilistic* closest-state function. We can identify a deterministic closest-state function \( f \) with the probabilistic closest-state function \( f' \) such that \( f'(\omega, i, \sigma'_i)(f(\omega, i, \sigma'_i)) = 1 \). With this identification, it should be clear that this definition of \( \mathcal{PR}_{i,\sigma'_i} \) is indeed a generalization of our earlier definition. Having made this change, the semantics of the formulas \( \text{RAT}_i \), \( B^*_i \varphi \), and \( \text{CB}^*_i \varphi \) remains unchanged, since they just use \( \mathcal{PR}_{i,\sigma'_i} \) as a black box.

Recall that in the informal discussion of how we could use deterministic closest-state functions in our original motivating example, we expanded the state space so that the states had all the information needed to remove uncertainty about the closest state. This intuition generalizes. The following result makes precise our claim that the use of deterministic closest-state functions is without loss of generality.

**Proposition 2.2** If \( M = (\Omega, s, f, \mathcal{PR}_1, \ldots, \mathcal{PR}_n) \) is a counterfactual structure with a probabilistic closest-state function \( f \), then there exists a counterfactual structure \( M' = (\Omega', s', f', \mathcal{PR}'_1, \ldots, \mathcal{PR}'_n) \) with a deterministic closest-state function \( f' \) and a surjective mapping \( G : \Omega' \to \Omega \) such that for all formulas \( \varphi \in \mathcal{L} \), we have

---

\(^8\) We can also show that \( \text{SRAT}^{k+1} \) implies \( \text{SRAT}^k \), but it is not the case that \( \text{SRAT}^k_i \) implies \( B^*_i \text{SRAT}^k_i \), since \( \text{RAT} \) does not imply \( B^*_i \text{RAT} \).

\(^9\) The results of this section are not used elsewhere in the paper. The reader who believes this claim can skip this section without loss of continuity.
\[(M', \omega') \models \varphi \text{ iff } (M, G(\omega')) \models \varphi.\]

Moreover, if \(M\) is strongly appropriate then so is \(M'\).

### 3 Characterizing common counterfactual belief of rationality

To put our result into context, we first restate the characterizations of rationalizability given by Tan and Werlang (1988) and Brandenburger and Dekel (1987) in our language. We first recall the definition of rationalizability given by Osborne and Rubinstein (1994):

**Definition 3.1** A strategy \(\sigma_i\) for player \(i\) is rationalizable if, for each player \(j\), there is a set \(Z_j \subseteq \Sigma_j(\Gamma)\) and, for each strategy \(\sigma_j' \in Z_j\), a probability measure \(\mu_{\sigma_j'}\) on \(\Sigma_{-j}(\Gamma)\) whose support is a subset of \(Z_{-j} = \prod_{j' \neq j} Z_{j'}\) such that

- \(\sigma_i \in Z_i\); and
- for strategy \(\sigma_j' \in Z_j\), strategy \(\sigma_j'\) is a best response to (the beliefs) \(\mu_{\sigma_j'}\).

A strategy profile \(\bar{\sigma}\) is rationalizable if every strategy \(\sigma_i\) in the profile is rationalizable. \(\square\)

**Theorem 3.2** (Brandenburger and Dekel 1987; Tan and Werlang 1988) \(\bar{\sigma}\) is rationalizable in a game \(\Gamma\) iff there exists a finite probability structure \(M\) that is appropriate for \(\Gamma\) and a state \(\omega\) such that \((M, \omega) \models \text{play}(\bar{\sigma}) \wedge \text{CB}(\text{RAT})\).

We now consider counterfactual structures. We here provide a condition on the closest-state function under which common (counterfactual) belief of rationality characterizes rationalizable strategies.

### 3.1 Counterfactual structures respecting unilateral deviations

Let \(M = (\Omega, f, \mathcal{PR}_1, \ldots, \mathcal{PR}_n)\) be a finite counterfactual structure that is appropriate for \(\Gamma\). \(M\) respects unilateral deviations if, for every state \(\omega \in \Omega\), player \(i\), and strategy \(\sigma_i'\) for player \(i\), \(s_{-i}(f(\omega, i, \sigma_i')) = s_{-i}(\omega)\) and \(\mathcal{PR}_{-i}(f(\omega, i, \sigma_i')) = \mathcal{PR}_{-i}(\omega)\); that is, in the closest state to \(\omega\) where player \(i\) switches strategies, everybody else’s strategy and beliefs remain same.

Recall that \(L^0\) is defined exactly like \(L\) except that we exclude the “counterfactual” modal operators \(B^*\) and \(CB^*\). The following theorem shows that for formulas in \(L^0\), counterfactual structures respecting unilateral deviations behave just as (standard) probability structures.

**Proposition 3.3** For every \(\varphi \in L^0\), there exists a finite probability structure \(M\) appropriate for \(\Gamma\) and a state \(\omega\) such that \((M, \omega) \models \varphi\) iff there exists a finite counterfactual structure \(M'\) (strongly) appropriate for \(\Gamma\) that respects unilateral deviations, and a state \(\omega'\) such that \((M', \omega') \models \varphi\).

\(^{10}\) This definition is not quite equivalent to the original definition of rationalizability due to Bernheim (1984) and Pearce (1984) (since it allows for opponents’ strategies to be correlated, where as Bernheim and Pearce require them to be independent).
We can now use Proposition 3.3 together with the standard characterization of common belief of rationality (Theorem 3.2) to characterize both common belief of rationality and common counterfactual belief of rationality.

**Theorem 3.4** The following are equivalent:

(a) \( \vec{\sigma} \) is rationalizable in \( \Gamma \);

(b) there exists a finite counterfactual structure \( M \) that is appropriate for \( \Gamma \) and respects unilateral deviations, and a state \( \omega \) such that \( (M, \omega) \models \text{play}(\vec{\sigma}) \land_{i=1}^{n} \text{WRAT}_k^i \) for all \( k \geq 0 \);

(c) there exists a finite counterfactual structure \( M \) that is strongly appropriate for \( \Gamma \) and respects unilateral deviations and a state \( \omega \) such that \( (M, \omega) \models \text{play}(\vec{\sigma}) \land_{i=1}^{n} \text{WRAT}_k^i \) for all \( k \geq 0 \);

(d) there exists a finite counterfactual structure \( M \) that is appropriate for \( \Gamma \) and respects unilateral deviations and a state \( \omega \) such that \( (M, \omega) \models \text{play}(\vec{\sigma}) \land_{i=1}^{n} \text{SRAT}_k^i \) for all \( k \geq 0 \);

(e) there exists a finite counterfactual structure \( M \) that is strongly appropriate for \( \Gamma \) and respects unilateral deviations and a state \( \omega \) such that \( (M, \omega) \models \text{play}(\vec{\sigma}) \land_{i=1}^{n} \text{SRAT}_k^i \) for all \( k \geq 0 \).

**Remark 3.5** Note that, in the proofs of Theorems 3.3 and 3.4, a weaker condition on the counterfactual structure would suffice, namely, that we restrict to counterfactual structures where, for every state \( \omega \in \Omega \), player \( i \), and strategy \( \sigma_i' \) for player \( i \), the projection of \( \mathcal{PR}_{i,\sigma_i'}(\omega) \) onto strategies and beliefs of players \(-i\) is equal to the projection of \( \mathcal{PR}_i(\omega) \) onto strategies and beliefs of players \(-i\). That is, every player’s counterfactual beliefs regarding other players’ strategies and beliefs are the same as the player’s actual beliefs.

### 3.2 Iterated minimax domination

We now characterize common counterfactual belief of rationality without putting any restrictions on the counterfactual structures (other than them being appropriate, or strongly appropriate). Our characterization is based on ideas that come from the characterization of rationalizability. It is well known that rationalizability can be characterized in terms of an iterated deletion procedure, where at each stage, a strategy \( \sigma \) for player \( i \) is deleted if there are no beliefs that \( i \) could have about the undeleted strategies for the players other than \( i \) that would make \( \sigma \) rational (Pearce 1984). Thus, there is a deletion procedure that, when applied repeatedly, results in only the rationalizable strategies, that is, the strategies that are played in states where there is common belief of rationality, being left undeleted. We now show that there is an analogous way of characterizing common counterfactual belief of rationality.

The key to our characterization is the notion of minimax-dominated strategies.

**Definition 3.6** Strategy \( \sigma_i \) for player \( i \) in game \( \Gamma \) is minimax dominated with respect to \( \Sigma_{-i} \subseteq \Sigma_{-i}(\Gamma) \) iff there exists a strategy \( \sigma_i' \in \Sigma_i(\Gamma) \) such that

\[
\min_{\tau_{-i} \in \Sigma_{-i}} u_i(\sigma_i', \tau_{-i}) > \max_{\tau_{-i} \in \Sigma_{-i}} u_i(\sigma_i, \tau_{-i}).
\]
In other words, player $i$’s strategy $\sigma_i$ is minimax dominated with respect to $\Sigma'_{-i}$ iff there exists a strategy $\sigma'_i$ such that the worst-case payoff for player $i$ if he uses $\sigma'_i$ is strictly better than his best-case payoff if he uses $\sigma_i$, given that the other players are restricted to using a strategy in $\Sigma'_{-i}$.

In the standard setting, if a strategy $\sigma_i$ for player $i$ is dominated by $\sigma'_i$ then we would expect that a rational player will never play $\sigma_i$, because $\sigma'_i$ is a strictly better choice. As is well known, if $\sigma_i$ is dominated by $\sigma'_i$, then there are no beliefs that $i$ could have regarding the strategies used by the other players according to which $\sigma_i$ is a best response (Pearce 1984). This is no longer the case in our setting. For example, in the standard setting, cooperation is dominated by defection in Prisoner’s Dilemma. But in our setting, suppose that player 1 believes that if he cooperates, then the other player will cooperate, while if he defects, then the other player will defect. Then cooperation is not dominated by defection.

So when can we guarantee that playing a strategy is irrational in our setting? This is the case only if the strategy is minimax dominated. If $\sigma_i$ is minimax dominated by $\sigma'_i$, there are no counterfactual beliefs that $i$ could have that would justify playing $\sigma_i$. Conversely, if $\sigma_i$ is not minimax dominated by any strategy, then there are beliefs and counterfactual beliefs that $i$ could have that would justify playing $\sigma_i$. Specifically, $i$ could believe that the players in $-i$ are playing the strategy profile that gives $i$ the best possible utility when he plays $\sigma_i$, and that if he switches to another strategy $\sigma'_i$, the other players will play the strategy profile that gives $i$ the worst possible utility given that he is playing $\sigma'_i$.

Note that we consider only domination by pure strategies. It is easy to construct examples of strategies that are not minimax dominated by any pure strategy, but are minimax dominated by a mixed strategy. Our characterization works only if we restrict to domination by pure strategies. The characterization, just as with the characterization of rationalizability, involves iterated deletion, but now we do not delete dominated strategies in the standard sense, but minimax-dominated strategies.

**Definition 3.7** Define $NSD_j^k(\Gamma)$ inductively: let $NSD_j^0(\Gamma) = \Sigma_j$ and let $NSD_j^{k+1}(\Gamma)$ consist of the strategies in $NSD_j^k(\Gamma)$ not minimax dominated with respect to $NSD_{-j}^k(\Gamma)$. Strategy $\sigma_i$ for player $i$ survives $k$ rounds of iterated deletion of minimax-dominated strategies for player $i$ if $\sigma_i \in NSD_i^k(\Gamma)$. Strategy $\sigma_i$ for player $i$ survives iterated deletion of minimax-dominated strategies if it survives $k$ rounds of iterated deletion of strongly dominated for all $k$, that is, if $\sigma_i \in NSD_i^{\infty}(\Gamma) = \cap_k NSD_i^k(\Gamma)$.

In the deletion procedure above, at each step we remove all strategies that are minimax dominated; that is, we perform a “maximal” deletion at each step. This was not necessary. The strategies that survives iterated deletion is actually independent of the deletion order. This result actually follows from a result of Gilboa et al. (1990), since, in their language, the ordering imposed by minimax dominance is hereditary and satisfies individual independent of irrelevant alternatives. Rather than explaining these notions, we provide a short self-contained proof of the result here.

Let $S^0, \ldots, S^m$ be sets of strategy profiles. $\hat{S} = (S^0, S^1, \ldots, S^m)$ is a terminating deletion sequence for $\Gamma$ if, for $j = 0, \ldots, m - 1$, $S^{j+1} \subset S^j$ (note that we use $\subset$ to mean proper subset) and all players $i$, $S_i^{j+1}$ contains all strategies for player $i$ not
minimax dominated with respect to $S_{-i}^j$ (but may also contain some strategies that are minimax dominated with respect to $S_{-i}^l$), and $S_i^m$ does not contain any strategies that are minimax dominated with respect to $S_{-i}^m$ (as well as no strategies minimax dominated with respect to $S_{-i}^{m-1}$). A set $T$ of strategy profiles has ambiguous terminating sets if there exist two terminating deletion sequences $\bar{S} = (T, S_1, \ldots, S_m)$, $\bar{S}' = (T, S'_1, \ldots, S'_m)$ such that $S_m \neq S'_m$; otherwise, we say that $T$ has a unique terminating set.

**Proposition 3.8** No (finite) set of strategy profiles has ambiguous terminating sets.

**Corollary 3.9** The set of strategies that survives iterated deletion of minimax-dominated strategies is independent of the deletion order.

**Remark 3.10** Note that in the definition of $\text{NSD}_i^k(\Gamma)$, we remove all strategies that are dominated by some strategy in $\Sigma_i(\Gamma)$, not just those dominated by some strategy in $\text{NSD}_i^{k-1}(\Gamma)$. Nevertheless, the definition would be equivalent even if we had considered only dominating strategies in $\text{NSD}_i^{k-1}(\Gamma)$. For suppose not. Let $k$ be the smallest integer such that there exists some strategy $\sigma_i \in \text{NSD}_i^{k-1}(\Gamma)$ that is minimax dominated by a strategy $\sigma_i' \notin \text{NSD}_i^{k-1}(\Gamma)$, but there is no strategy in $\text{NSD}_i^{k-1}(\Gamma)$ that dominates $\sigma_i$. That is, $\sigma_i'$ was deleted in some previous iteration. Then there exists a sequence of strategies $\sigma_i^0, \ldots, \sigma_i^q$ and indices $k_0 < k_1 < \ldots < k_q = k - 1$ such that $\sigma_i^0 = \sigma_i', \sigma_i^j \in \text{NSD}_i^{k_j}(\Gamma)$, and for all $0 \leq j < q$, $\sigma_i^j$ is minimax dominated by $\sigma_i^{j+1}$ with respect to $\text{NSD}_i^{k_j}(\Gamma)$. Since $\text{NSD}_i^{k-2}(\Gamma) \subseteq \text{NSD}_i^{k-j}(\Gamma)$ for $j \leq k - 2$, an easy induction on $j$ shows that $\sigma_i^q$ minimax dominates $\sigma_i^{q-j}$ with respect to $\text{NSD}_i^{k-2}$ for all $0 < j \leq q$. In particular, $\sigma_i^q$ minimax dominates $\sigma_i^0 = \sigma_i'$ with respect to $\text{NSD}_i^{k-2}$.

The following example shows that iteration has bite: there exist a 2-player game where each player has $k$ actions and $k - 1$ rounds of iterations are needed.

**Example 3.11** Consider a two-player game, where both players announce a value between 1 and $k$. Both players receive in utility the smallest of the values announced; additionally, the player who announces the larger value gets a reward of $p = 0.5$. That is, $u(x, y) = (y + p, y)$ if $x > y$, $(x, x + p)$ if $y > x$, and $(x, x)$ if $x = y$. In the first step of the deletion process, 1 is deleted for both players; playing 1 can yield a maximum utility of 1, whereas the minimum utility of any other action is 1.5. Once 1 is deleted, 2 is deleted for both players: 2 can yield a max utility of 2, and the minimum utility of any other action (once 1 is deleted) is 2.5. Continuing this process, we see that only $(k, k)$ survives.

**3.3 Characterizing iterated minimax domination**

We now show that strategies surviving iterated removal of minimax-dominated strategies characterize the set of strategies consistent with common counterfactual belief.

---

11 This game can be viewed as the “opposite” of Traveler’s Dilemma (Basu 1994), where the player who announces the smaller value gets the reward.
of rationality in (strongly) appropriate counterfactual structures. As a first step, we define a “minimax” analogue of rationalizability.

**Definition 3.12** A strategy $\sigma_i$ for player $i$ in game $\Gamma$ is **minimax rationalizable** if, for each player $j$, there is a set $Z_j \subseteq \Sigma_j(\Gamma)$ such that

1. $\sigma_i \in Z_i$;
2. for each player $j$, strategy $\sigma'_j \in Z_j$, and strategy $\sigma''_j \in \Sigma_j(\Gamma)$,

$$\max_{\tau_{-j} \in \mathcal{Z}_{-j}} u_j(\sigma'_j, \tau_{-j}) \geq \min_{\tau_{-j} \in \mathcal{Z}_{-j}} u_j(\sigma''_j, \tau_{-j}).$$

A strategy profile $\vec{\sigma}$ is minimax rationalizable if each strategy $\sigma_i$ in the profile is minimax rationalizable.

**Theorem 3.13** The following are equivalent:

(a) $\vec{\sigma} \in \text{NSD}^\infty(\Gamma)$;

(b) $\vec{\sigma}$ is minimax rationalizable in $\Gamma$;

(c) there exists a finite counterfactual structure $M$ that is strongly appropriate for $\Gamma$ and a state $\omega$ such that $(M, \omega) \models \text{play}(\vec{\sigma}) \land \bigwedge_{i=1}^n \text{SRAT}_i^k$ for all $k \geq 0$;

(d) for all players $i$, there exists a finite counterfactual structure $M$ that is appropriate for $\Gamma$ and a state $\omega$ such that $(M, \omega) \models \text{play}_i(\sigma_i) \land \text{SRAT}_i^k$ for all $k \geq 0$.

### 4 Characterizing analogues of Nash equilibrium

In this section, we consider analogues of Nash equilibrium in our setting. This allows us to relate our approach to the work of Tennenholz (2004) and Kalai et al. (2010). In the standard setting, if a strategy profile $\vec{\sigma}$ is a Nash equilibrium, then there exists a state where $\vec{\sigma}$ is played, common belief of rationality holds, and additionally, the strategy profile is (commonly) known to the players (Aumann and Brandenburger 1995). To study analogues of Nash equilibrium, we thus investigate the effect of adding assumptions about knowledge of the players’ strategies. Define the event $\text{KS}$ as follows:

1. $(M, \omega) \models \text{KS}$ iff, for all players $i$,

$$\mathcal{PR}_i(\omega)([s_{-i}(\omega)]_M) = 1.$$

That is KS holds if every player $i$ knows the strategies used by players $-i$. We can now define a **translucent equilibrium** as a strategy profile $\vec{\sigma}$ that can be played in a state where both KS and common counterfactual belief of rationality holds:

**Definition 4.1** A pure strategy profile $\vec{\sigma}$ is a **translucent equilibrium** if there exists a finite counterfactual structure $M$ that is appropriate for $\Gamma$ and a state $\omega$ such that $(M, \omega) \models \text{KS} \land \text{play}(\vec{\sigma}) \land \bigwedge_{i=1}^n \text{SRAT}_i^k$ for every $k \geq 0$.  

__Capraro and Halpern (2015) generalize this definition to allow mixed strategy profiles. Allowing mixed strategies would distract from the issues that we want to focus on in this section, so we have considered only translucent equilibria in pure strategies.__
KS does not require that player \( i \) knows how players \(-i\) will respond to \( i \) switching strategies. However, as we now show, the notion of translucent equilibrium is quite robust. Making stronger assumptions about the players’s knowledge would not affect it. One stronger condition would be to require not only that every player \( i \) knows the strategies of the other players, but also how they respond to \( i \) switching strategies.

- \((M, \omega) \models \text{KR iff, for all players } i \text{ and strategies } \sigma'_i \text{ for } i,\)

\[
P^cR_{i,\sigma'_i}(\omega)([s_{-i}(f(\omega, i, \sigma'_i))]_M) = 1.
\]

Clearly, KR implies KS (by simply considering \( \sigma'_i = s_i(\omega) \)). An even stronger condition is to require that the players know the true state of the world.

- \((M, \omega) \models \text{KW iff, for all players } i,\)

\[
P^cR_i(\omega)(\omega) = 1.
\]

If all players know the true state of the world, then they also counterfactually know the true state of the world: for every player \( i \) and every strategy \( \sigma'_i \) for player \( i \),

\[
P^cR_{i,\sigma'_i}(\omega)(f(\omega, i, \sigma'_i)) = 1.
\]

It follows that KW implies KR and thus also KS. Additionally, note that KW implies \( EB(KW) \), so KW also implies \( CB(KW) \).

We now characterize CCBR in structures satisfying the conditions above. As we show, the characterization is the same, no matter which “knowledge of strategy” condition we use (and thus we could have equivalently defined translucent equilibria using any of them). To do this, we first need to recall the definition of individual rationality.

**Definition 4.2** A strategy profile \( \tilde{\sigma} \) is *individually rational* (IR) if, for every player \( i \) in the game \( \Gamma \),

\[
u_i(\tilde{\sigma}) \geq \max_{\sigma'_i \in \Sigma_i(\Gamma)} \min_{\tau_{-i} \in \Sigma_{-i}(\Gamma)} u_i(\sigma'_i, \tau_{-i}).
\]

Although every IR strategy profile is contained in \( NSD^1(\Gamma) \), it is not necessarily contained in \( NSD^2(\Gamma) \). That is, IR strategies may not survive two rounds of deletion of minimax-dominated strategies. To see this, consider the game \( \Gamma \) in Example 3.11. Both players’ maximin payoff is 1.5, so every strategy profile in \( NSD^1(\Gamma) = \{(x, y) \mid 2 \leq x, y \leq k\} \) is IR, but \( NSD^2(\Gamma) \) does not contain \((2, 2)\).

As the following simple example shows, not every strategy profile that survives iterated deletion of minimax-dominated strategies is IR.
Example 4.3 Consider the game with payoffs given in the table below.

|   | c       | d       |
|---|---------|---------|
| a | (100, 0)| (100, 0)|
| b | (150, 0)| (50, 0) |

All strategy profiles survive iterated deletion of minimax-dominated strategies, but \((b, d)\) is not individually rational since playing \(a\) always guarantees the row player utility 100. 

Let \(IR(\Gamma)\) denote the set of IR strategy profiles in \(\Gamma\), and let \(IR(\mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n, \Gamma) = IR(\Gamma')\) where \(\Gamma'\) is the subgame of \(\Gamma\) obtained by restricting player \(i\)'s strategy set to \(\mathcal{Z}_i\). That is, \(IR(\mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n, \Gamma)\) is the set of strategies \(\bar{\sigma} \in \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n\) such that for every player \(i\),

\[
\forall \bar{\sigma}, \sigma_i' \in \mathcal{Z}_i, \tau_{-i} \in \mathcal{Z}_{-i}, \quad u_i(\bar{\sigma}) \geq \min_{\sigma_i' \in \mathcal{Z}_i} \max_{\tau_{-i} \in \mathcal{Z}_{-i}} u_i(\sigma_i', \tau_{-i}).
\]

A stronger way of capturing individual rationality of subgames is to require that the condition above hold even if the max is taken over every \(\sigma_i' \in \Sigma(\Gamma)\) (as opposed to only \(\sigma_i' \in \mathcal{Z}_i\)). More precisely, let \(IR'(\mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n, \Gamma)\) be the set of strategies \(\bar{\sigma} \in \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n\) such that, for all players \(i\),

\[
\forall \bar{\sigma}, \sigma_i' \in \Sigma(\Gamma), \tau_{-i} \in \mathcal{Z}_{-i}, \quad u_i(\bar{\sigma}) \geq \min_{\sigma_i' \in \Sigma(\Gamma)} \max_{\tau_{-i} \in \mathcal{Z}_{-i}} u_i(\sigma_i', \tau_{-i}).
\]

We now state our characterization of translucent equilibria.

**Theorem 4.4** The following are equivalent:

(a) \(\bar{\sigma} \in IR(\text{NSD}^\infty(\Gamma), \Gamma)\);
(b) \(\bar{\sigma} \in IR'(\text{NSD}^\infty(\Gamma), \Gamma)\);
(c) \(\bar{\sigma}\) is minimax rationalizable and \(\bar{\sigma} \in IR'(\mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n, \Gamma)\), where \(\mathcal{Z}_1, \ldots, \mathcal{Z}_n\) are the sets of strategies guaranteed to exists by the definition of minimax rationalizability;
(d) there exists a finite counterfactual structure \(M\) that is strongly appropriate for \(\Gamma\) and a state \(\omega\) such that \((M, \omega) \models KW \land \text{play}(\bar{\sigma}) \land \bigwedge_{i=1}^n \text{SRAT}^k_i\) for every \(k \geq 0\);
(e) \(\bar{\sigma}\) is a translucent equilibrium; that is, there exists a finite counterfactual structure \(M\) that is appropriate for \(\Gamma\) and a state \(\omega\) such that \((M, \omega) \models KS \land \text{play}(\bar{\sigma}) \land \bigwedge_{i=1}^n \text{SRAT}^k_i\) for every \(k \geq 0\).

As a consequence of Theorem 4.4, Theorem 3.13, and Example 4.3, we have that whereas every translucent equilibrium is minimax rationalizable, not every minimax rationalizable strategy is a translucent equilibrium.

It is worth comparing Theorem 4.4 to the results of Tennenholtz (2004) and Kalai et al. (2010) on program equilibria/equilibria with conditional commitments. Recall that these papers focus on 2-player games. In Tennenholtz’s model, each player \(i\)
deterministically picks a program $\Pi_i$; player $i$’s action is $\Pi_i(\Pi_{-i})$. In the two-player case, a program equilibrium is a pair of programs $(\Pi_1, \Pi_2)$ such that no player can improve its utility by unilaterally changing its program. In this model any IR strategy profile $(a_1, a_2)$ can be sustained in a program equilibrium: each player uses the program $\Pi$, where $\Pi(\Pi')$ outputs $a_i$ if $\Pi' = \Pi$, and otherwise “punishes” the other player using his minmax strategy. (Tennenholtz extends this result to show that any mixed IR strategy profile can be sustained in a program equilibrium, by considering randomizing programs; Kalai et al. show that all correlated IR strategy profiles can be sustained, by allowing the players to pick a distribution over programs.) In contrast, in our model, a smaller set of strategy profiles can be sustained. This difference can be explained as follows. In the program equilibrium model a player may “punish” the other player using an arbitrary action (e.g., using minimax punishment) although this may be detrimental for him. Common counterfactual belief of rationality disallows such punishments. More precisely, it allows a player $i$ to punish other players only by using a strategy that is rational for player $i$. On the other hand, as we now show, if we require only common belief (as opposed to counterfactual belief) in rationality, then any IR strategy can be sustained in an equilibrium in our model.

**Theorem 4.5** The following are equivalent:

(a) $\vec{\sigma} \in IR(\Gamma)$;
(b) there exists a finite counterfactual structure $M$ that is strongly appropriate for $\Gamma$ and a state $\omega$ such that $(M, \omega) \models KW \land play(\vec{\sigma}) \land CB(RAT)$;
(c) there exists a finite counterfactual structure $M$ that is appropriate for $\Gamma$ and a state $\omega$ such that $(M, \omega) \models KS \land play(\vec{\sigma}) \land CB(RAT)$.

**5 Discussion**

We have introduced a game-theoretic framework for analyzing scenarios where a player may believe that if he were to switch strategies, this might affect the distribution of other players’ strategy profiles that he faces. We considered solution concepts for this setting that generalize rationalizability and Nash equilibrium; we leave it to future work to explore other solution concepts.

As we pointed out in the introduction, the setting that we consider seems to arise not just in human interactions, but also more broadly in interactions between species. Capraro and Halpern (2015) have shown that thinking in terms of translucency can help explain observed behavior in social dilemmas, games with a unique Nash equilibrium, and a unique social welfare-maximizing strategy profile that gives each player a higher utility than they get in the Nash equilibrium. Games such as Prisoner’s Dilemma, Traveler’s Dilemma (Basu 1994), the public goods game, and Bertrand competition) are all instances of social dilemmas. Experiments have considered how the degree of cooperation (i.e., the extent to which the welfare-maximizing strategy is played) in social dilemmas varies with parameters such as the payoffs and the number of players. Capraro and Halpern showed that if we assume that, if a player deviates from cooperation, the other players will realize it with probability $\alpha$ and then also defect (i.e., play their component of the Nash equilibrium), and make a few other (quite
minimal and natural) assumptions, then we can explain all the regularities observed in the literature on social dilemmas. Capraro and Halpern also show how translucency to a “God” can help explain pro-social behavior even in single-agent games (i.e., decision problems). Recent work (Johnson 2016; Norenzayan 2013) has argued that belief that “God is watching” is quite ubiquitous. This suggests that thinking in terms of translucency might be a useful tool when it comes to analyzing human behavior.

Our formal model allows players’ counterfactual beliefs (i.e., their beliefs about the state of the world in the event that they switch strategies) to be arbitrary—they may be completely different from the players’ actual beliefs. It would be of great interest to investigate some restrictions on players counterfactual beliefs. Capraro and Halpern already did this in their setting. Another natural restriction would be to require that a player $i$’s counterfactual beliefs regarding other players’ strategies and beliefs be $\epsilon$-close to player $i$’s actual beliefs in total variation distance\(^{13}\) —that is, for every state $\omega \in \Omega$, player $i$, and strategy $\sigma_i'$ for player $i$, the projection of $\mathcal{PR}^c_{1,\sigma_i}(\omega)$ onto strategies and beliefs of players $-i$ is $\epsilon$-close to the projection of $\mathcal{PR}_i(\omega)$ onto strategies and beliefs of players $-i$. We refer to counterfactual structures satisfying this property as $\epsilon$-counterfactual structures. Roughly speaking, $\epsilon$-counterfactual structures restrict to scenarios where players are not “too” transparent to one another; this captures the situation when a player assigns only some “small” probability to its switch in action being noticed by the other players. 0-counterfactual structures behave just as counterfactual structures that respect unilateral deviations: common counterfactual belief of rationality in 0-counterfactual structures characterizes rationalizable strategies (see Remark 3.5). The general counterfactual structures investigated in this paper are 1-counterfactual structures (that is, we do not impose any conditions on players’ counterfactual beliefs). We remark that although our characterization results rely on the fact that we consider 1-counterfactual structures, the motivating example in the introduction (the translucent Prisoner’s Dilemma) shows that even considering $\epsilon$-counterfactual structures with a small $\epsilon$ can result in there being strategies consistent with common counterfactual belief of rationality that are not rationalizable. We leave a fuller exploration of common counterfactual belief of rationality in $\epsilon$-counterfactual structures for future work.

Another way of restricting players’ counterfactual beliefs is to consider various dynamics that describe how other players’ strategies might change as a consequence of $i$ changing his strategy. That is, we can consider various ways that a “chain reaction” (in the spirit of the discussion in the introduction) might proceed. Are there natural assumptions that could be made about such a process?

More generally, it seems to us that thinking about translucency should open up a host of interesting avenues for further exploration.

---

\(^{13}\) Recall that two probability distribution are $\epsilon$-close in total variation distance if the probabilities that they assign to any event $E$ differ by at most $\epsilon$. 

\(\text{Springer}\)
A Proofs

In this appendix, we prove the results whose proofs were omitted in the main text. We repeat the statements of the results for the reader’s convenience.

**Proposition 2.2** If \( M = (\Omega, s, f, PR_1, \ldots, PR_n) \) is a counterfactual structure with a probabilistic closest-state function \( f \), then there exists a counterfactual structure \( M' = (\Omega', s', f', PR'_1, \ldots, PR'_n) \) with a deterministic closest-state function \( f' \) and a surjective mapping \( G : \Omega' \rightarrow \Omega \) such that for all formulas \( \varphi \in L \), we have

\[
(M', \omega') \models \varphi \iff (M, G(\omega')) \models \varphi.
\]

Moreover, if \( M \) is strongly appropriate then so is \( M' \).

**Proof** We change \( f \) player by player, state by state, and strategy by strategy. More precisely, we construct a sequence \( M^h = (\Omega^h, s^h, f^h, PR^h_1, \ldots, PR^h_n) \) of counterfactual structures and surjective mappings \( G^h : \Omega^h \rightarrow \Omega \) for \( h = 0, \ldots, |\Omega|(|\Sigma_1| + \cdots + |\Sigma_n|) \) such that (a) for all formulas \( \varphi \in L \), we have \( (M^h, \omega') \models \varphi \iff (M, G^h(\omega')) \models \varphi 
\), (b) \( M^0 = M \), and (c) there exist \( h \) distinct tuples \((i, \omega, \sigma_i)\) with \( \omega \in \Omega \) and \( \sigma_i \in \Sigma_i \) such that if \( G^h(\omega') = \omega \), then \( f^h(i, \omega', \sigma_i) \) is deterministic. Clearly, \( f^{|\Omega|(|\Sigma_1|+\cdots+|\Sigma_n|)} \) must be deterministic, so \( M^{|\Omega|(|\Sigma_1|+\cdots+|\Sigma_n|)} \) is the desired structure \( M' \). We proceed as follows.

Set \( M^0 = M \). Suppose that we have constructed \( M^0, \ldots, M^h \) and \( h < |\Omega|(|\Sigma_1| + \cdots + |\Sigma_n|) \). If \( f^h \) is deterministic, we can take \( M^{h+1} = M^h \). If not, there is some tuple \((i, \omega', \sigma_i)\) such that \( f^h(i, \omega', \sigma_i) \) is not deterministic. Suppose that \( G^h(\omega') = \omega \) and the support of \( f^h(i, \omega', \sigma_i) = \{\omega^1, \ldots, \omega^N\} \). Note that \( s(\omega^j) = \sigma_i \) for \( j = 1, \ldots, N \) and \( \omega^j \neq \sigma_i \) (for otherwise \( f(i, \omega', \sigma_i) \) would be \( \omega' \), and \( f \) would be deterministic.) Let \( \Omega^h+1 \) be the result of adding \( N - 1 \) new states, call them \( \omega^2, \ldots, \omega^N \), to \( \Omega^h \) for each state \( \omega^* \in \Omega^h \) such that \( G^h(\omega^*) = \omega \). (Implicitly, we identify \( \omega^* \) with \( \omega^* \) in the construction below.) Define \( f^{h+1}(i, \omega^*, \sigma_I) = \omega^\ell \) for \( \ell = 1, \ldots, N \). (Since we are identifying \( \omega^* \) with \( \omega^* \), \( f^{h+1}(i, \omega^*, \sigma_I) = \omega^1 \).) Thus, \( f^{h+1}(i, \omega^*, \sigma_I) \) is deterministic. Define \( f^{h+1}(j, \omega^*, \sigma_J) = f^h(j, \omega^*, \sigma_J) \) if \( i \neq j \) or \( \sigma_I \neq \sigma_J \). Finally, \( f^{h+1}(j, \omega', \sigma_J) = f^h(j, \omega', \sigma_J) \) for all players \( j \) and strategies \( \sigma_J \) if \( G(\omega') \neq \omega \). Thus, \( f^{h+1} \) agrees with \( f^h \) except on inputs of the form \((i, \omega^*, \sigma_J)\). Define \( G^{h+1}(\omega') = G^h(\omega') \) if \( G^h(\omega') \neq \omega \) and \( G^{h+1}(\omega^*) = G^h(\omega^*) \) for all the states \( \omega^* \in \Omega^{h+1} - \Omega^h \). Thus, \( f^{h+1} \) is deterministic on all the tuples for which \( f^h \) is deterministic, and, in addition, is deterministic on tuples \((i, \omega^*, \sigma_I)\) such that \( G^{h+1}(\omega^*) = \omega \). Define \( s^{h+1}(\omega') = s(G(\omega')) \) for all \( \omega' \in \Omega^{h+1} \). Finally, define \( PR_j^{h+1}(\omega')(\omega'') = PR_j^h(\omega')(\omega'')(\omega') \) for all players \( j \) if \( G^h(\omega'') \neq \omega \) and \( \omega'' \in \Omega^h \); \( PR_{h+1}(\omega')(\omega'')(\omega') = (PR_j^h(\omega')(\omega'')(\omega'))(f^h(i, \omega', \sigma_J)) \); and \( PR_{h+1}(\omega') = PR_{h+1}(\omega'') \) for \( \ell = 2, \ldots, N \). Let \( M^{h+1} = (\Omega^{h+1}, f^{h+1}, PR_1^{h+1}, \ldots, PR_n^{h+1}) \).

With these definitions, we leave it to the reader to check that \( (M^h, \omega') \models \varphi \iff (M, G^h(\omega')) \models \varphi \), and that \( M^h \) is strongly appropriate if \( M \) is. This completes the proof. \( \square \)
Proposition 3.3 For every $\varphi \in \mathcal{L}^0$, there exists a finite probability structure $M$ appropriate for $\Gamma$ and a state $\omega$ such that $(M, \omega) \models \varphi$ iff there exists a finite counterfactual structure $M'$ (strongly) appropriate for $\Gamma$ that respects unilateral deviations, and a state $\omega'$ such that $(M', \omega') \models \varphi$.

Proof For the “if” direction, let $M' = (\Omega, f, PR_1, \ldots, PR_n)$ be a finite counterfactual structure that is counterfactually appropriate for $\Gamma$ (but not necessarily strongly counterfactually appropriate) and respects unilateral deviations. Define $M = (\Omega, PR_1, \ldots, PR_n)$. Clearly $M$ is a finite probability structure appropriate for $\Gamma$; it follows by a straightforward induction on the length of $\varphi$ that $(M', \omega) \models \varphi$ iff $(M, \omega) \models \varphi$.

For the “only-if” direction, let $M = (\Omega, PR_1, \ldots, PR_n)$ be a finite probability structure, and let $\omega \in \Omega$ be a state such that $(M, \omega) \models \varphi$. We assume without loss of generality that for each strategy profile $\vec{\sigma}'$ there exists some state $\omega_{\vec{\sigma}'} \in \Omega$ such that $s(\omega_{\vec{\sigma}'} = \vec{\sigma}'$ and for each player $i$, $PR_i(\omega_{\vec{\sigma}'})(\omega_{\vec{\sigma}'} = 1$. (If such a state does not exist, we can always add it.)

We define a finite counterfactual structure $M' = (\Omega', f', PR'_{R_1}, \ldots, PR'_{R_n})$ as follows:

- $\Omega' = \{ (\vec{\sigma}', \omega') : \vec{\sigma}' \in \Sigma(\Gamma), \omega' \in \Omega \}$;
- $s'(\vec{\sigma}', \omega') = \vec{\sigma}'$;
- $f((\vec{\sigma}', \omega'))i, \sigma'' \) = ((\sigma^i, \sigma''_i), \omega')$;
- $PR'_{R_i}(\omega')(\omega'', \omega') = \begin{cases} PR_i(\omega')(\omega'') & \text{if } \sigma' = s(\omega'), \sigma'' = s(\omega'') \\ 1 & \text{if } \sigma' \neq s(\omega'), (\sigma'', \omega'') = (\sigma', \omega_{\vec{\sigma}'}) \\ 0 & \text{otherwise.} \end{cases}$

It follows by construction that $M'$ is strongly appropriate for $\Gamma$ and respects unilateral deviations. Furthermore, it follows by an easy induction on the length of the formula $\varphi'$ that for every state $\omega \in \Omega$, $(M, \omega) \models \varphi'$ iff $(M', (s(\omega), \omega)) \models \varphi'$.

Theorem 3.4 The following are equivalent:

(a) $\vec{\sigma}$ is rationalizable in $\Gamma$;
(b) there exists a finite counterfactual structure $M$ that is appropriate for $\Gamma$ and respects unilateral deviations, and a state $\omega$ such that $(M, \omega) \models \text{play}(\vec{\sigma}) \wedge_{i=1}^n \text{WRAT}^k_i$ for all $k \geq 0$;
(c) there exists a finite counterfactual structure $M$ that is strongly appropriate for $\Gamma$ and respects unilateral deviations and a state $\omega$ such that all $k \geq 0$;
(d) there exists a finite counterfactual structure $M$ that is appropriate for $\Gamma$ and respects unilateral deviations and a state $\omega$ such that $(M, \omega) \models \text{play}(\vec{\sigma}) \wedge_{i=1}^n \text{SRAT}^k_i$ for all $k \geq 0$;
(e) there exists a finite counterfactual structure $M$ that is strongly appropriate for $\Gamma$ and respects unilateral deviations and a state $\omega$ such that $(M, \omega) \models \text{play}(\vec{\sigma}) \wedge_{i=1}^n \text{SRAT}^k_i$ for all $k \geq 0$.

Proof The equivalence of (a), (b), and (c) is immediate from Theorem 3.2, Theorem 3.3, and Proposition 2.1. We now prove the equivalence of (b) and (d). Consider an counterfactual structure $M$ that is appropriate for $\Gamma$ and respects unilateral deviations. The result follows immediately once we show that for all states $\omega$ and all
for all players $i$, there exists a finite counterfactual structure $M$ that is appropriate for $\Gamma_i$. Thus, for all $\omega' \in \text{Supp}(P^R_i(\omega))$, we have that $(M, \omega') \models \land_{j \neq i} \text{WRAT}_{j}^{k-1}$. Thus, by the induction hypothesis, $(M, \omega') \models \land_{j \neq i} \text{SRAT}_{j}^{k-1}$. Since, as we have observed, the truth of a formula of the form $B_j^i \varphi$ at a state $\omega''$ depends only on $j$’s beliefs at $\omega''$ and the truth of $\text{RAT}_j$ depends only on $j$’s strategy and beliefs at $\omega''$, it easily follows that, if $j$ has the same beliefs and plays the same strategy at $\omega_1$ and $\omega_2$, then $(M, \omega_1) \models \text{SRAT}_{j}^{k-1} \iff (M, \omega_2) \models \text{SRAT}_{j}^{k-1}$. Since $(M, \omega') \models \land_{j \neq i} \text{SRAT}_{j}^{k-1}$ and $M$ respect unilateral deviations, for all strategies $\sigma_i'$, it follows that $(M, f(\omega', i, \sigma_i')) \models \land_{j \neq i} \text{SRAT}_{j}^{k-1}$. Thus, $(M, \omega) \models \text{RAT}_i \land B_j^i (\land_{j \neq i} \text{SRAT}_{j}^{k-1})$, as desired. The argument that $(c)$ is equivalent to $(e)$ is identical; we just need to consider strongly appropriate counterfactual structures rather than just appropriate counterfactual structures. \hfill $\Box$

**Proposition 3.8** No (finite) set of strategy profiles has ambiguous terminating sets.

**Proof** Let $T$ be a set of strategy profiles of least cardinality that has ambiguous terminating deletion sequences $\vec{S} = (T, S_1, \ldots, S_m)$ and $\vec{S}' = (T', S_1', \ldots, S_m')$, where $S_m \neq S_m'$. Let $T'$ be the set of strategies that are not minimax dominated with respect to $T$. Clearly $T' \neq \emptyset$ and, by definition, $T' \subseteq S_1 \cap S_1'$. Since $T'$, $S_1$, and $S_1'$ all have cardinality less than that of $T$, they must all have unique terminating sets; moreover, the terminating sets must be the same. For consider a terminating deletion sequence starting at $T'$. We can get a terminating deletion sequence starting at $S_1$ by just appending this sequence to $S_1$ (or taking this sequence itself, if $S_1 = T'$). We can similarly get a terminating deletion sequence starting at $S_1'$. Since all these terminating deletion sequences have the same final element, this must be the unique terminating set. But $(S_1, \ldots, S_m)$ and $(S_1', \ldots, S_m')$ are terminating deletion sequences with $S_m \neq S_m'$, a contradiction. \hfill $\Box$

**Theorem 3.13** The following are equivalent:

(a) $\vec{\sigma} \in \text{NSD}^\infty(\Gamma)$;
(b) $\vec{\sigma}$ is minimax rationalizable in $\Gamma$;
(c) there exists a finite counterfactual structure $M$ that is strongly appropriate for $\Gamma$ and a state $\omega$ such that all $k \geq 0$;
(d) for all players $i$, there exists a finite counterfactual structure $M$ that is appropriate for $\Gamma$ and a state $\omega$ such that $(M, \omega) \models \text{play}_i(\sigma_i) \land \text{SRAT}_i^k$ for all $k \geq 0$.

**Proof** We prove that (a) implies (b) implies (c) implies (d) implies (a). We first introduce some helpful notation. Recall that $\arg \max_x g(x) = \{y : \text{for all } z, g(z) = g(y)\}$; $\arg \min_x g(x)$ is defined similarly. For us, $x$ ranges over pure strategies or pure strategy profiles, and we will typically be interested in considering some element of the set, rather than the whole set. Which element we take does not matter. For definiteness, we define a tie-breaking rule by assuming that there is some order on the set of pure
strategies and strategy profiles, and take the arg max\textsuperscript{x} g (x) to be the maximum element of arg max\textsubscript{x} g (x) with respect to this order; arg min\textsuperscript{x} g (x) is defined similarly.

(a) \Rightarrow (b): Let K be an integer such that NSD\textsuperscript{K} (\Gamma) = NSD\textsuperscript{K+1} (\Gamma); such a K must exist since the game is finite. It also easily follows that for each player j, NSD\textsuperscript{K} (\Gamma) is non-empty: in iteration k + 1, no NSD\textsuperscript{K} -maximim strategy, that is, no strategy in arg max\textsubscript{\sigma_j' \in NSD\textsuperscript{K} (\Gamma)} min\textsubscript{\tau_j \in NSD\textsuperscript{K} (\Gamma)} u_j (\sigma_j', \tau_j), is deleted, since no maximin strategy is minimax dominated by a strategy in NSD\textsuperscript{K} (\Gamma) (recall that by Remark 3.10, it suffices to consider domination by strategies in NSD\textsuperscript{K} (\Gamma)). Let Z' = NSD\textsuperscript{K} (\Gamma). It immediately follows that the sets Z', ..., Z'\textsubscript{n} satisfy the conditions of Definition 3.12.

(b) \Rightarrow (c): Suppose that \tilde{\sigma} is minimax rationalizable. Let Z, ..., Z\textsubscript{n} be the sets guaranteed to exist by Definition 3.12. Let W\textsuperscript{d} = \{ (\tilde{\sigma}, i) | \tilde{\sigma} \in Z - i \times \Sigma_i \}, and let W\textsuperscript{0} = \{ (\tilde{\sigma}, 0) | \tilde{\sigma} \in Z_1 \times \ldots \times Z_n \}. Think of W\textsuperscript{0} as states where common counterfactual belief of rationality holds, and of W\textsuperscript{d} as “counterfactual” states where player i has changed strategies. In states in W\textsuperscript{0}, each player j assigns probability 1 to the other players choosing actions that maximize j’s utility (given his action). On the other hand, in states in W\textsuperscript{d}, where i ≠ 0, player i assigns probability 1 to the other players choosing actions that minimize i’s utility, whereas all other player j ≠ i still assign probability 1 to other players choosing actions that maximize j’s utility.

Define a structure M = (Ω, f, s, PR\textsubscript{1}, ..., PR\textsubscript{n}), where

- Ω = \bigcup_{i \in \{0,1,\ldots,n\}} W\textsuperscript{d};
- s(\tilde{\sigma}', i) = \tilde{\sigma}'
- PR\textsubscript{j}(\tilde{\sigma}', i)(\tilde{\sigma}'', i') = \begin{cases} 1 & \text{if } i = j = i', \sigma''_j = \sigma'_j, \text{ and } \sigma''_{-j} = \arg \min_{\tau_{-j} \in Z_{-j}} u_j (\sigma'_j, \tau_{-j}), \\ 0 & \text{if } i ≠ j, i' ≠ 0, \text{ and } \sigma''_j = \sigma'_j, \text{ and } \sigma''_{-j} = \arg \max_{\tau_{-j} \in Z_{-j}} u_j (\sigma'_j, \tau_{-j}), \end{cases}
- f((\tilde{\sigma}', i), j, \sigma''_j) = \begin{cases} (\tilde{\sigma}', i) & \text{if } \sigma''_j = \sigma'_j, \\ ((\sigma''_j, \tau_{-j}), j) & \text{otherwise}, \text{ where } \tau_{-j} = \arg \min_{\tau_{-j} \in Z_{-j}} u_j (\sigma'_j, \tau_{-j}). \end{cases}

It follows by inspection that M is strongly appropriate for \Gamma. We now prove by induction on k that, for all k ≥ 1 all i ∈ {0,1, ..., n}, and all states \omega ∈ W\textsuperscript{d}, (M, \omega) \models \wedge_{j ≠ i} SRAT\textsuperscript{k}.

For the base case (k = 1), since SRAT\textsuperscript{1} is logically equivalent to RAT\textsubscript{j}, we must show that if \omega ∈ W\textsuperscript{d}, then (M, \omega) \models \wedge_{j ≠ i} RAT\textsubscript{j}. Suppose that \omega = (\tilde{\sigma}', i) ∈ W\textsuperscript{d}. If i ≠ j, then at \omega, player j places probability 1 on the true state being \omega' = (\tilde{\sigma}'', 0), where \sigma''_j = \sigma'_j and \sigma''_{-j} = \arg \max_{\tau_{-j} \in Z_{-j}} u_j (\sigma'_j, \tau_{-j}). Player j must be rational, since if there exists some strategy \tau'_j such that u_j (\tilde{\sigma}'') < \sum_{\omega'' \in Ω} PR\textsubscript{j}(\omega'')(\omega'' | u_j (\tau'_j, s_{-j} (\omega'))), then the definition of PR\textsubscript{j} guarantees that u_j (\tilde{\sigma}'') < u_j (\tau'_j, \tau''_{-j}), where \tau''_j = \arg \min_{\tau_{-j} \in Z_{-j}} u_j (\sigma'_j, \tau_{-j}). If this inequality holds, then \tau'_j would minimax dominate \sigma'_j, contradicting the assumption that \sigma'_j ∈ Z_{-j}. 
For the induction step, suppose that the result holds for \( k \); we show that it holds for \( k + 1 \).
Suppose that \( \omega \in \cal W^j \) and \( j \neq i \). By construction, the support of \( \cal PR_j(\omega) \) is a subset of \( \cal W^j \); by the induction hypothesis, it follows that \( (\cal M, \omega) \models \cal R_j(\bigwedge_{j' \neq j} \cal SRAT^k_{j'}) \). Moreover, by construction, it follows that for all players \( j \) and all strategies \( \sigma'_j \neq s_i(\omega) \), the support of \( \cal PR_{j,\sigma'_j}(\omega) \) is a subset of \( \cal W^j \). By the induction hypothesis, it follows that for all \( j \neq i \), \( \omega \models B_j^*(\bigwedge_{j' \neq j} \cal SRAT^k_{j'}) \). Finally, it follows from the induction hypothesis that for all \( j \neq i \), \( \cal R_j(\bigwedge_{j' \neq j} \cal SRAT^k_{j'}) \). Since \( \cal SRAT^k_{j'} \) implies \( \cal R_j \), it follows that for all \( j \neq i \), \( \omega \models \cal R_j \land B_j^*(\bigwedge_{j' \neq j} \cal SRAT^k_{j'}) \), which proves the induction step.

(c) \( \Rightarrow \) (d): The implication is trivial.

(d) \( \Rightarrow \) (a): We prove an even stronger statement: For all \( k \geq 0 \), if there exists a finite counterfactual structure \( \cal M^k \) that is appropriate for \( \Gamma \) and a state \( \omega \) such that \( (\cal M^k, \omega) \models \text{play}_i(\sigma_i) \land \cal SRAT_i^k \), then \( \sigma_i \in \cal NSD_k^i(\Gamma) \). 14 We proceed by induction on \( k \).

The result clearly holds if \( k = 0 \). Suppose that the result holds for \( k - 1 \) for \( k \geq 1 \); we show that it holds for \( k \). Let \( \cal M^k = (\Omega, f, s, \cal P_1, \ldots, \cal P_n) \) be a finite counterfactual structure that is appropriate for \( \Gamma \) and a state \( \omega \) such that \( (\cal M^k, \omega') \models \text{play}_i(\sigma_i) \land \cal SRAT_i^k \). Replacing \( \cal SRAT_i^k \) by its definition, we get that

\[
(\cal M^k, \omega') \models \text{play}_i(\sigma_i) \land \cal R_i^* \land \cal B_i^*(\bigwedge_{j \neq i} \cal SRAT_{j}^{k-1}).
\]

By definition of \( \cal B_i^* \), it follows that for all strategies \( \sigma'_i \) for player \( i \) and all \( \omega'' \) such that \( \cal PR_{i,\sigma'_i}(\omega'')(\omega'') > 0 \),

\[
(\cal M^k, \omega'') \models \bigwedge_{j \neq i} \cal SRAT_{j}^{k-1},
\]

so by the induction hypothesis, it follows that for all \( \omega'' \) such that \( \cal PR_{i,\sigma'_i}(\omega'')(\omega'') > 0 \), we have \( s_{-i}(\omega'') \in \cal NSD_{i-1}^{k'}(\Gamma) \). Since \( (\cal M^k, \omega') \models \text{play}_i(\sigma_i) \land \cal R_i \), it follows that \( \sigma_i \) cannot be minimax dominated with respect to \( \cal NSD_{i-1}^{k'}(\Gamma) \). Since, for all \( j' > 1 \), \( \cal NSD_{i-1}^{j'}(\Gamma) \subseteq \cal NSD_{i-1}^{j'-1}(\Gamma) \), it follows that, for all \( k' < k \), \( \sigma_i \) is not minimax dominated with respect to \( \cal NSD_i^{k'}(\Gamma) \). Thus, \( \sigma_i \in \cal NSD_i^k(\Gamma) \).

**Theorem 4.4** The following are equivalent:

(a) \( \tilde{\sigma} \in IR(\cal NSD^\infty_i(\Gamma), \Gamma) \);
(b) \( \tilde{\sigma} \in IR(\cal NSD^\infty_i(\Gamma), \Gamma) \);
(c) \( \tilde{\sigma} \) is minimax rationalizable and \( \tilde{\sigma} \in IR(\cal I, \cal Z_1 \times \cdots \times \cal Z_n, \Gamma) \), where \( \cal Z_1, \ldots, \cal Z_n \) are the sets of strategies guaranteed to exists by the definition of minimax rationalizability;
(d) there exists a finite counterfactual structure \( \cal M \) that is strongly appropriate for \( \Gamma \) and a state \( \omega \) such that \( (\cal M, \omega) \models \text{KW} \land \text{play}(\tilde{\sigma}) \land \bigwedge_{i=1}^n \cal SRAT_i^k \) for every \( k \geq 0 \).

14 The converse also holds; we omit the details.
(e) \( \vec{\sigma} \) is a translucent equilibrium; that is, there exists a finite counterfactual structure \( M \) that is appropriate for \( \Gamma \) and a state \( \omega \) such that \( (M, \omega) \models KS \land \text{play}(\vec{\sigma}) \land \bigwedge_{i=1}^{n} \text{SRAT}_{k}^{i} \) for every \( k \geq 0 \).

Proof. Again, we prove that (a) implies (b) implies (c) implies (d) implies (e) implies (a).

\((a) \Rightarrow (b)\): We show that if \( \vec{\sigma} \in IR(\text{NSD}^{K}(\Gamma), \Gamma) \) then \( \vec{\sigma} \in IR'(\text{NSD}^{K}(\Gamma), \Gamma) \). The implication then follows from the fact that since the game is finite there exists some \( K \) such that \( \text{NSD}^{K}(\Gamma) = \text{NSD}^{\infty}(\Gamma) \).

Assume by way of contradiction that \( \vec{\sigma} \in IR(\text{NSD}^{K}(\Gamma), \Gamma) \) but \( \vec{\sigma} \notin IR'(\text{NSD}^{K}(\Gamma), \Gamma) \); that is, there exists a player \( i \) and a strategy \( \sigma_{i}' \notin \text{NSD}^{K}_{i}(\Gamma) \) such that

\[
\min_{\tau_{-i} \in \text{NSD}^{K}_{-i}(\Gamma)} u_{i}(\sigma_{i}', \tau_{-i}) > u_{i}(\vec{\sigma}).
\]

By the argument in Remark 3.10, there exists a strategy \( \sigma_{i}'' \in \text{NSD}^{K}_{i}(\Gamma) \) such that \( u_{i}(\sigma_{i}'', \tau_{-i}) > u_{i}(\sigma_{i}', \tau_{-i}') \) for all \( \tau_{-i}' \in \text{NSD}^{K}_{-i}(\Gamma) \). It follows that

\[
\min_{\tau_{-i} \in \text{NSD}^{K}_{-i}(\Gamma)} u_{i}(\sigma_{i}'', \tau_{-i}) > u_{i}(\vec{\sigma}).
\]

Thus, \( \vec{\sigma} \notin IR(\text{NSD}^{K}(\Gamma), \Gamma) \).

\((b) \Rightarrow (c)\): The implication follows in exactly the same way as in the proof that (a) implies (b) in Theorem 3.13.

\((c) \Rightarrow (d)\): Suppose that \( \vec{\sigma} \) is minimax rationalizable. Let \( Z_{1}, \ldots, Z_{n} \) be the sets guaranteed to exist by Definition 3.12, and suppose that \( \vec{\sigma} \in IR'(Z_{1} \times Z_{n}, \Gamma) \). Define the sets \( W^{i} \) as in the proof of Theorem 3.13. Define the structure \( M \) just as in the proof of Theorem 3.13, except that for all players \( i \), let \( \mathcal{P}R_{i}((\vec{\sigma}, 0))((\vec{\sigma}'', i')) = 1 \) in case \( \vec{\sigma}' = \vec{\sigma} \) and \( i' = 0 \). Clearly \( (M, (\vec{\sigma}, 0)) \models KW \). It follows using the same arguments as in the proof of Theorem 3.13 that \( M \) is strongly appropriate and that \( (M, (\vec{\sigma}, 0)) \models \text{play}(\vec{\sigma}) \land \bigwedge_{i=1}^{n} \text{SRAT}_{k}^{i} \) for every \( k \geq 0 \); we just need to rely on the (strong) IR property of \( \vec{\sigma} \) to prove the base case of the induction.

\((d) \Rightarrow (e)\): The implication is trivial.

\((e) \Rightarrow (a)\): Recall that since the game is finite, there exists a constant \( K \) such that \( \text{NSD}^{K-1}(\Gamma) = \text{NSD}^{K}(\Gamma) = \text{NSD}^{\infty}(\Gamma) \). We show that if there exists a finite counterfactual structure \( M \) that is appropriate for \( \Gamma \) and a state \( \omega \) such that \( (M, \omega) \models KS \land \text{play}(\vec{\sigma}) \land \bigwedge_{i=1}^{n} \text{SRAT}_{k}^{i} \), then \( \vec{\sigma} \in IR(\text{NSD}^{K}(\Gamma), \Gamma) \).

Consider some state \( \omega \) such that \( (M, \omega) \models KS \land \text{play}(\vec{\sigma}) \land \bigwedge_{i=1}^{n} \text{SRAT}_{k}^{i} \). By Theorem 3.13, it follows that \( \vec{\sigma} \in \text{NSD}^{K}(\Gamma) \). For each player \( i \), it additionally follows that \( (M, \omega) \models \text{play}(\vec{\sigma}) \land \text{EB}(\text{play}(\vec{\sigma})) \land \text{RAT}_{i} \land B_{i}^{\ast}(\bigwedge_{j \neq i} \text{SRAT}_{k}^{j-1}) \). By Theorem 3.13, it follows that for every strategy \( \sigma_{i}' \) for \( i \), the support of the projection of \( \mathcal{P}R_{i, \sigma_{i}'}^{c}(\omega) \)
onto strategies for players $-i$ is a subset of $\text{NSD}_{-i}^K(\Gamma) = \text{NSD}_{-i}^K(\Gamma)$. Thus, we have that for every $\sigma_i'$, there exists $\tau_{-i} \in \text{NSD}_{-i}^K(\Gamma)$ such that $u_i(\tilde{\sigma}) \geq u_i(\sigma_i', \tau_{-i})$, which means that $\tilde{\sigma}$ is IR in the subgame induced by restricting the strategy set to $\text{NSD}^K(\Gamma)$.

$\square$

**Theorem 4.5** The following are equivalent:

(a) $\tilde{\sigma} \in \text{IR}(\Gamma)$;

(b) there exists a finite counterfactual structure $M$ that is strongly appropriate for $\Gamma$ and a state $\omega$ such that $(M, \omega) \models \text{KW} \land \text{play}(\tilde{\sigma}) \land \text{CB}(\text{RAT})$;

(c) there exists a finite counterfactual structure $M$ that is appropriate for $\Gamma$ and a state $\omega$ such that $(M, \omega) \models \text{KS} \land \text{play}(\tilde{\sigma}) \land \text{CB}(\text{RAT})$.

**Proof** Again, we prove that (a) implies (b) implies (c) implies (a).

\[(a) \Rightarrow (b):\] Define a structure $M = (\Omega, f, s, \mathcal{P}R_1, \ldots, \mathcal{P}R_n)$, where

- $\Omega = \Sigma(\Gamma)$;
- $s(\tilde{\sigma}') = \tilde{\sigma}'$;
- $\mathcal{P}R_j(\tilde{\sigma}')(\tilde{\sigma}') = 1$;
- $f(\tilde{\sigma}', i, \sigma_j') = \begin{cases} \tilde{\sigma}' & \text{if } \sigma_j' = \sigma_j'' \\ (\sigma_j'', \tau_{-j}') & \text{otherwise}, \text{ where } \tau_{-j}' = \arg\min_{\tau_{-j} \in \Sigma_{-j}(\Gamma)} u_j(\sigma_j', \tau_{-j}). \end{cases}$

It follows that $M$ is strongly appropriate for $\Gamma$ and that $(M, \tilde{\sigma}) \models \text{KW}$. Moreover, $(M, \tilde{\sigma}) \models \text{RAT}$ since $\tilde{\sigma}$ is individually rational; furthermore, since each player considers only the state $\tilde{\sigma}$ possible at $\tilde{\sigma}$, it follows that $(M, \tilde{\sigma}) \models \text{CB}(\text{RAT})$.

\[(b) \Rightarrow (c):\] The implication is trivial.

\[(c) \Rightarrow (a):\] Suppose that $M = (\Omega, f, s, \mathcal{P}R_1, \ldots, \mathcal{P}R_n)$ is a finite counterfactual structure appropriate for $\Gamma$, and $(M, \omega) \models \text{KW} \land \text{play}(\tilde{\sigma}) \land \text{CB}(\text{RAT})$. It follows that for each player $i$, $(M, \omega) \models \text{play}(\tilde{\sigma}) \land \text{EB}(\text{play}(\tilde{\sigma})) \land \text{RAT}_i$. Thus, we have that for all strategies $\sigma_i'$, there exists $\tau_{-i} \in \Sigma_{-i}(\Gamma)$ such that $u_i(\tilde{\sigma}) \geq u_i(\sigma_i', \tau_{-i})$, which means that $\tilde{\sigma}$ is IR.

$\square$

**References**

Aumann, R. J. (1995). Backwards induction and common knowledge of rationality. *Games and Economic Behavior*, 8, 6–19.

Aumann, R. J., & Brandenburger, A. (1995). Epistemic conditions for Nash equilibrium. *Econometrica*, 63(5), 1161–1180.

Basu, K. (1994). The traveler’s dilemma: paradoxes of rationality in game theory. *The American Economic Review*, 84(2), 391–395.

Bernheim, B. D. (1984). Rationalizable strategic behavior. *Econometrica*, 52(4), 1007–1028.

Brandenburger, A., & Dekel, E. (1987). Rationalizability and correlated equilibria. *Econometrica*, 55, 1391–1402.

Capraro, V. (2013). A solution concept for games with altruism and cooperation. arXiv:1302.3988.

Capraro, V., Halpern, J.Y. (2015). Translucent players: explaining cooperative behavior in social dilemmas. Proceedings of the 15th conference on Theoretical Aspects of Rationality and Knowledge, 2015. https://ssrn.com/abstract=2509678.
Daley, B., Sadowski, P. (2014). A strategic model of magical thinking: axioms and analysis. http://www.princeton.edu/economics/seminar-schedule-by-prog/behavioralf14/Daley_Sadowski_MT.pdf.
Ekman, P., & Friesen, W. (1969). Nonverbal leakage and clues to deception. Psychiatry, 32, 88–105.
Eshel, I., & Cavalli-Sforza, L. L. (1982). Assortment of encounters and evolution of cooperativeness. Proceedings of the National Academy of Sciences, 79, 1331–1335.
Gilboa, I., Kalai, E., & Zemel, E. (1990). On the order of eliminating dominated strategies. Operations Research Letters, 9, 85–89.
Halpern, J. Y. (1999). Hypothetical knowledge and counterfactual reasoning. International Journal of Game Theory, 28(3), 315–330.
Howard, N. (1971). Paradoxes of rationality: theory of metagames and political behavior. Cambridge: MIT Press.
Johnson, D. (2016). God is watching you: how the fear of god makes us human. Oxford: Oxford University Press.
Kalai, A., Kalai, E., Lehrer, E., & Samet, D. (2010). A commitment folk theorem. Games and Economic Behavior, 69(1), 127–137.
Kalay, A. (2003). Friends or foes? Empirical test of a one-period division game having a unique Nash equilibrium. unpublished manuscript.
Lewis, D. (1973). Counterfactuals. Cambridge: Harvard University Press.
Masel, J. (2007). A Bayesian model of quasi-magical thinking can explain observed cooperation in the public good game. Journal of Economic Behavior and Organization, 64(1), 216–231.
Norenzayan, A. (2013). Big gods: how religion transformed cooperation and conflict. Princeton: Princeton University Press.
Osborne, M. J., & Rubinstein, A. (1994). A course in game theory. Cambridge: MIT Press.
Pearce, D. G. (1984). Rationalizable strategic behavior and the problem of perfection. Econometrica, 52(4), 1029–1050.
Salcedo, B. (2013). Implementation without commitment in moral hazard environments. arXiv:1303.0916.
Samet, D. (1996). Hypothetical knowledge and games with perfect information. Games and Economic Behavior, 17, 230–251.
Shafir, E., & Tversky, A. (1992). Thinking through uncertainty: Nonconsequential reasoning and choice. Cognitive Psychology, 24, 449–474.
Solan, E., & Yariv, L. (2004). Games with espionage. Games and Economic Behavior, 47, 172–199.
Spohn, W. (2003). Dependency equilibria and the causal structure of decision and game situations. Homo Oeconomicus, 20, 195–255.
Stalnaker, R. C. (1968). A theory of conditionals. In N. Rescher (Ed.), Studies in logical theory (pp. 98–112). Oxford: Blackwell.
Stalnaker, R. C. (1996). Knowledge, belief and counterfactual reasoning in games. Economics and Philosophy, 12, 133–163.
Tan, T., & Werlang, S. (1988). The Bayesian foundation of solution concepts of games. Journal of Economic Theory, 45(45), 370–391.
Tennenholtz, M. (2004). Program equilibrium. Games and Economic Behavior, 49(12), 363–373.
Tillio, A. D., Halpern, J. Y., & Samet, D. (2014). Conditional belief types. Games and Economic Behavior, 87, 253–268.
von Neumann, J., & Morgenstern, O. (1947). Theory of games and economic (2nd ed.). Princeton: Princeton University Press.
Zambrano, E. (2004). Counterfactual reasoning and common knowledge of rationality in normal form games. Topics in Theoretical Economics., https://doi.org/10.2202/1534-598X.1020.