In this paper, we present a model of rationality in extensive form games in which players, during the course of the game, may revise their conjectures about the opponents’ preferences, while imposing common belief of rationality at each information set. Since players are assumed to be expected utility maximizers, their preferences, at each of their information sets, are given by a utility function over the reachable terminal nodes and some subjective probability distribution over the opponents’ strategy choices compatible with this information set. A player may therefore revise his conjecture about an opponent’s utility function, or his subjective probability distribution, or both, whenever this opponent has chosen a move which would be suboptimal given the previously conjectured preferences.

We refer to these two types of revisions as utility revisions and probability revisions, respectively. A profile of conjectures about the opponents’ preferences, as described above, is required to be self-enforcing in the sense that, if these conjectures were to be made public, no player, at any of his information sets, would have an incentive to change his conjecture about the opponents’ preferences. More precisely, if player $i$ believes that player $j$, at information set $h$, assigns positive probability to some player $k$ strategy, then this strategy should be optimal for player $k$, given player $j$’s conjecture at $h$ about player $k$’s preferences. We refer to this property as sequential rationality.

In addition, we impose a restriction on the revision of conjectures about the opponents’ utility functions, called utility consistency, which states that a player, who at information set $h$ decides to revise his conjecture about an opponent’s utility function, should do this in such a way that the opponent’s utilities following $h$ remain unaltered. A last property we require is that, at the beginning of the game, the players hold common conjectures about the other players’ utility functions and, for every player $i$, there exists a common initial conjecture about the initial subjective probability distribution attached to player $i$’s strategy choice. A profile of conjectures satisfying the latter property, together with sequential rationality and utility consistency, shall be called a preference conjecture equilibrium.

It is shown that, for every game tree and every possible profile $u$ of utility functions at the terminal nodes, there exists a preference conjecture equilibrium in which the players’ initial common conjectures about the players’ utility
functions coincide with \( u \). This result thus establishes that, within this framework, common belief of rationality at every information set is always possible for every given profile of initial common conjectures about the players’ utility functions. However, there may exist preference conjecture equilibria that could be rejected on the ground that they contain “too many” utility and/or probability revisions. Intuitively, one would wish for a theory of rationality in which the number of preference conjecture revisions, needed to maintain common belief of rationality throughout the game, is minimal. We will refer to this property as the minimum revision principle.

As to the question how to measure preference conjecture revisions, we adopt the view that utility revisions should have a higher weight than probability revisions. The reason is that a player’s utility function is, in some sense, a more “robust” characteristic for this player than his subjective probability distribution about the opponents’ strategies since the latter depends crucially upon his perception of the other players’ preferences, whereas the former does not. Formalizing this point of view leads to the concept of minimum revision equilibrium, selecting those preference conjecture equilibria for which the set of utility revisions is minimal, and for which there is no other equilibrium with the same set of utility revisions but fewer probability revisions. The forward induction flavour of minimum revision equilibrium is being illustrated by its performance in several examples to be discussed in this paper. In particular, we prove that the concept of minimum revision equilibrium uniquely selects the forward induction outcome in every outside option game, as defined in van Damme (1989). We show that in perfect information games with generic utility functions \( u \), every preference conjecture equilibrium for which the initial common conjectures about the players’ utility functions coincide with \( u \), always supports the backward induction procedure in this game.