September 30, 2009
Eric Rasmusen, erasmuse@indiana.edu

A Paradox of Sequential Rationality

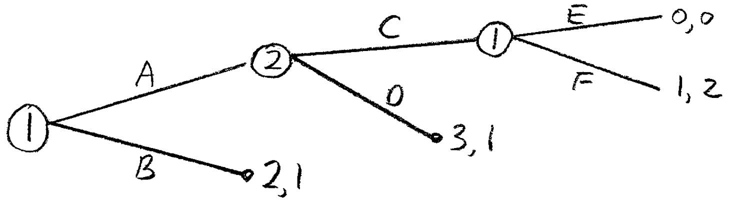
Consider the perfect-information game in the figure below. Player 1's strategy set is: $(A, E|AC), (A, F|AC), (B, E|AC), (B, F|AC)$. Player 2's strategy set is: $C|A, D|A$.

Using backwards induction, Player 1 uses $F|AC$, Player 2 uses $C|A$, and Player 1 uses B in the subgame perfect Nash equilibrium.

The other Nash equilibrium is for Player 1 to use $(A, E|AC)$ and for Player 2 to use $D|A$.

Player 1 prefers the non-perfect equilibrium with its (3,1) instead of (2,1) payoffs. The story justifying subgame perfectness that we tell ourselves is that before the game starts, if Player 1 threatens Player 2, saying: "I am going to choose A, so you'd better choose D in response, or I'll go ahead and choose E and we'll both get 0," then Player 2 will respond, "I don't believe you. I'm calling your bluff. You are rational, and so I know you would never choose E instead of F." Then Player 1 would give up and choose B instead.

But what if Player 1 makes his little speech, and then actually does choose A? What should Player 2 think? In equilibrium, choosing A isn't supposed to happen. It seems to refute the assumption of Player 1 being rational. So maybe Player 2 should respond by choosing D after all. If he does that, however, then Player 1's action has turned out to be rational after all.



(This is inspired by Section 6.4 of Osborne and Rubinstein's 1994 *A Course in Game Theory*, which has a similar but not identical game.)