DE GRUYTER OPEN

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 50 (63) 2017

DOI: 10.1515/slgr-2017-0016

Marek M. Kamiński
University of California, Irvine

BACKWARD INDUCTION: MERITS AND FLAWS

Abstract. Backward induction (BI) was one of the earliest methods developed for solving finite sequential games with perfect information. It proved to be especially useful in the context of Tom Schelling's ideas of credible versus incredible threats. BI can be also extended to solve complex games that include an infinite number of actions or an infinite number of periods. However, some more complex empirical or experimental predictions remain dramatically at odds with theoretical predictions obtained by BI. The primary example of such a troublesome game is Centipede. The problems appear in other long games with sufficiently complex structure. BI also shares the problems of subgame perfect equilibrium and fails to eliminate certain unreasonable Nash equilibria.

Keywords: backward induction, Nash equilibrium, subgame perfect equilibrium, sequential game, extensive form game, Centipede.

Introduction: the incredible importance of incredible threats

While the origins of backward induction (BI) are unclear, it owes its importance to the concept of incredible threat introduced by Schelling (1960). It will be helpful to start the discussion of BI with a few illustrative examples of credible and incredible threats coming from the Cold War, whose epic strategic dilemmas made game theory so popular.

The interplay between credible versus incredible threats was at the center of the hugely popular movie "Dr. Strangelove." The central idea in "Dr. Strangelove" was the establishment by the Soviet Union of a "Doomsday Machine," an automatic system that after a nuclear attack on Soviet territory would launch a devastating blast of 50 Cobalt-Thorium G bombs which in two months would encircle Earth and kill all life. Since the "Doomsday Machine" couldn't be stopped, it created a *credible threat*: any potential aggressor would anticipate that the response to his actions would be lethal

ISBN 978-83-7431-527-2 ISSN 0860-150X

and thus he would prefer not to attack. The twist in the movie was that the accidental communication problems between the two superpowers prevented the US from taking the doomsday response into calculation and contributed to the mistake that started a global nuclear apocalypse.

Credible threats (or credible commitments) appear both in everyday life, our language, and historical anecdotes. When Hernán Cortés burned his ships after landing in Mexico in February, 1519, he committed himself and his men to staying and fighting. There was no return. Spartacus killing his horse at Salerno or commanders burning bridges behind their armies similarly commit to fighting. Credible commitment – or a credible threat – means in this case eliminating the possibility of retreat.

In order to see the difference between a credible threat and an incredible one, let's consider a simple game representing the flavors of Cold War clashes of the 1950s and 1960s and, at the same time, the nightmarish dream of American military analysts (see Figure 1).

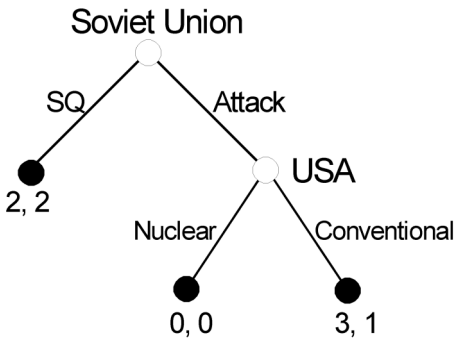

Figure 1. Cold War Threat

The story behind our example is as follows:

The Soviets contemplate spreading their revolutionary regime to the West and attacking Western Europe (Attack). If they attack, the United States can respond with Nuclear or Conventional forces. The Soviet conventional forces are much stronger than the American and Western European forces combined, and the result of conventional warfare would be a Soviet victory. The payoffs associated with the possible three outcomes represent the standard preferences that we can ascribe to the two players.

The Cold War Threat has two Nash equilibria. In the first NE, the Soviets do nothing since their attack would be met with an American nuclear strike. In the second NE, the Soviets attack and the Americans launch only a conventional defense since nuclear war would be too costly for all involved

parties. The problem with the first NE is that it is based on an *incredible threat* of Americans using nuclear weapons after being attacked. The threat is incredible since, after the Americans are attacked, they face a new reality and have to re-evaluate their options. They consider responding Nuclear (payoff 0) and Conventional (payoff 1). Since the payoff for conventional response is higher than the payoff for nuclear response, they do not have incentive to carry out their threat. Instead, they rationally choose the action that gives them a higher payoff.

The second NE is free from the problem of incredible threats.

This is the potential nightmare of American strategists: nuclear weapons may turn out to be useless since, when decisions have to be made, it will be more beneficial to refrain from using them. In the analysis identifying which NE is based on credible versus incredible threats, the critical moment comes when the US (the second player) considers the last move. The choice is made in favor of the action bringing higher payoff.

Backward induction cleverly disregards Nash equilibria that are based on incredible threats. In our example, BI rejects the Nuclear response as an equilibrium strategy versus Conventional since the former generates lower payoff when we compare payoffs from the American last moves. Thus, BI offers a *refinement* of Nash equilibrium, i.e., for every game, the solutions obtained by BI are a subset of all Nash equilibria. Excluded from consideration are equilibria that are intuitively "unreasonable."

Before we define BI in a more precise fashion, let's take a brief history tour. Zermelo (1913) was first to analyze winning in chess that is a natural application of backward induction, but his method of analysis was based on a different principle than BI (Schwalbe and Walker, 2001). Reasoning based on BI was implicit in Stackelberg's (1934) construction of his equilibrium alternative to Cournot (1838). As a general procedure for solving two-person zero-sum games of perfect information, BI appeared in von Neumann and Morgenstern's 1944 founding book (1944: 117). It was used to prove a precursor of Kuhn's Theorem for chess and similar games. Von Neumann's exceedingly complex formulation was later clarified and elevated to high theoretical status by Kuhn's work (1953, especially Corollary 1). While being very intuitive, Schelling's idea of "incredible threat" turned out to be difficult to formalize. Moreover, BI eliminates other unreasonable Nash equilibria that are not based on incredible threats. The NE refinement that was both defined formally and could also account for various sorts of unreasonable NE was offered by Selten's (1965) introduction of subgame perfection. In the late 1970s, BI came under critical fire when the chain-store paradox, Centipede, and other games questioned its univer-

sal virtue (Selten, 1978; Rosenthal, 1981). The criticism, supported by experimental evidence, greatly diminished the appeal of backward induction in some games. Backward induction is extensively discussed by Binmore (1987, 1988) and in popular textbooks by Fudenberg and Tirole (1991) and Myerson (1991).

Backward induction for finite games of perfect information

Let's describe the BI procedure in more detail and how it removes incredible threats by using two games that involve such incredible threats.

A (politically correct) bum approaches you in a dark street, displays a gun, and asks for a $100 donation. If you give him the money (M), the game ends. If you refuse his offer (R), he promises to shoot *himself* (see the left panel in Figure 2 BUM).

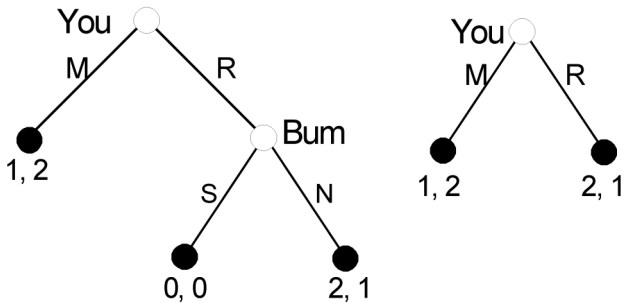

Figure 2. The game of Bum (left panel) and Bum reduced after applying the first step of backward induction (right panel)

The solution to the game based on backward induction would proceed as follows: First, let's consider what happens when you refuse to give the money to the bum. He chooses between shooting himself and receiving the payoff of 0 versus not shooting and receiving 1; thus, he selects N. In the first case, you receive 0 and in the second case, you receive 2. You can anticipate his action N and assume that refusing automatically gives you 2 (since he won't shoot himself) while giving the money to the bum gives you 1. You choose not to give him the money.

In your reasoning, you used a small part of the original game, where the bum's subgame is substituted with the payoffs resulting from his best choice (Figure 2, right panel). By analyzing the smaller game, you choose R, and the strategy profile chosen by BI is (R,N). The game was solved by

running backward with respect to the decision-making order. First, the last decision in the game was analyzed and then the reasoning turned to the first decision.

The main idea behind the BI algorithm is – by running backward in the game – to keep reducing games to smaller games, and to keep in mind the partial solution obtained in the process of reduction. Slightly more formally, the algorithm may be described for any finite extensive form game of perfect information in the following way:

Start at any of the final decision nodes, D, that is followed only by terminal nodes and let's assume that the player who moves at D is j. Then select an action that offers the highest payoff to j at D (let's call such an action a). The existence of at least one payoff-maximizing action is guaranteed by the finite character of the game. Let's assume that there is just a single action that leads to the highest payoff. (If there are many actions with such a property, the reasoning is conducted separately for all of them and the more general algorithm applicable to such a case is slightly more complex.) Next, a new game is constructed by substituting D and the branches following D in the old game with the payoff vector assigned to a. The intuition behind the reduction is that all players in the game may predict that, once D is reached, j will correctly choose the action a that maximizes his payoff. Thus, our procedure assumes that once the game reaches D, the payoff vector assigned to a will automatically follow. This means that there is no point in considering D and the following actions. The game may be now simplified. One only has to remember the action a that j selected at D since this particular action becomes a part of the game's solution.

The reasoning described above, i.e., the reduction of the original game to smaller games, is repeated until a final single vector of payoffs is obtained. What results from the reduction is a strategy profile that consists of all best actions obtained at some stage of reduction procedure. Those actions (including the actions that will not be actually played) constitute a backward induction strategy profile. It can be proved that, if there are no ties in payoffs at any stage of reduction, the final strategy profile is the same for all possible reductions of the original game.

When there are two or more actions available, one has to create separate partial strategy profiles for all such actions that include all previously obtained actions. Then the reasoning proceeds in a similar fashion for every partial strategy profile.

In the game of Bum, the strategy profile selected by backward induction was (R,N). It is easy to see that this particular strategy profile constitutes a Nash equilibrium of this game. If Player 1 switches from R to M, then her

payoff goes down from 2 to 1. If Player 2 switches from N to S, his payoff goes down from 1 to 0. This example illustrates an important property of BI: it selects only Nash equilibria.

(R,N) is not the only Nash equilibrium. Another one is (M,S) in which you offer the bum his donation; if you didn't, he would shoot himself. This equilibrium has one flaw: it doesn't look reasonable. It is based on an *incredible threat* that the bum would shoot himself once you refuse the donation. However, if you refuse, the bum has no incentive to carry out the threat. He would be better off by abandoning his threat. This NE based on incredible threat corresponds to the NE in our Cold War Threat game that involved American nuclear response to the Soviet attack.

Let's assume now that our bum has an option of shooting himself also when he receives the 100$ donation. Thus, the game does not end with the donation M but two more actions are available to the bum once he receives the donation: s and n (see Figure 3 Crazy Bum).

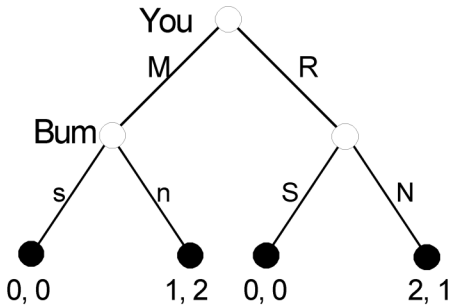

Figure 3. Crazy Bum

In Crazy Bum, there are three Nash equilibria. (M,nS) and (R,nN) intuitively correspond to (M,S) and (R,N) in the original Bum. The third equilibrium, (R,sN), is the weirdest one. In this equilibrium you refuse to make the donation, and the bum doesn't shoot himself. However, he would shoot himself if you gave him the money! The behavior associated with the strategy profile (R,sN) makes even less sense that the one associated with the equilibrium based on an incredible threat. And yet (R,sN) is a Nash equilibrium.

Thus, there are NE that look unreasonable but do not involve any incredible threats. While BI removes equilibria of the sort described above, the informal concept of incredible threat does not provide any justification for such removal. A more general and formal theory, that also encompassed the concept of incredible threat, was provided by Selten's concept of subgame perfect equilibrium.

Subgame perfection and variants of backward induction

The ideas incorporating incredible threats and other unreasonable equilibria were formalized by Selten (1965) with his solution concept of subgame perfect equilibrium. Informally, a strategy profile is SPE if it is a Nash equilibrium in all subgames of a game, including the game itself. The requirement imposed by the Nash equilibrium on players is strengthened in SPE. Since a subgame can be imagined as a small self-enclosed game, in an SPE, players are required to play Nash-equilibrium strategies at every decision node that starts such a new smaller, independently played game. In the game of Bum, SPE requires that the Nash equilibrium in this game must have an additional property: in the subgame when the bum makes his decision, he must choose an action that is better for him. This is precisely the requirement of eliminating incredible threats. In the game of Crazy Bum, SPE requires additionally that the bum must choose the best action after receiving the money. This means imposing a requirement that is not directly related to incredible threat.

It is relatively easy to prove that for finite games of perfect information, backward induction and subgame perfection always bring the same solutions, i.e., the SPE and the set of solutions obtained by BI are identical for all such games. It can be also proved that this coincidence is in fact true for all games (according to a very general definition of a game) and all pure strategies (Kaminski, 2009). Thus, the final important fact about BI is that it selects precisely the same set of strategy profiles as subgame perfection. At the same time, for many games it offers an effective algorithm of finding all SPEs that is simpler than looking for SPEs directly from its definition.

The more general algorithm for finding all SPEs, described in Kaminski (2009), is grounded on complicated axiomatic foundations and is too complex to define in its entirety here. Instead, I will describe below four finite and non-finite games that can be solved for SPEs using more general versions of the basic algorithm.

Pick-an-Integer: In the version of BI algorithm described earlier, the reduction of the game was obtained by removing a single set of branches that originated at the same decision node. This version of backward induction can be generalized and multiple branches can be removed simultaneously.

Consider the following two-player game: you start with choosing a positive integer that is not greater than ten. Next, your opponent chooses an-

other integer that is greater than your number, but the difference is again not greater than ten. Then you choose another integer by adding to the previous number no more than ten, and so on. The player who is the first to say a number at least equal to 100 is the winner.

Who can claim victory in this game?

The reader may be willing to pause reading at this point and to try solving the game. In fact, the problem is not easy to solve until we apply backward induction. Then it becomes trivial:

Since the game is finite, there is a winner in each specific play. Let's assume that the winning player made the total of n moves. The key observation is that she is guaranteed to say 100 only if she said 89 in her move $n-1$. If she says less than 89 in her move $n-1$, then her opponent could say 89 or less, and she would be able to say no more than 99. If she says 90 or more, her opponent could win by saying 100. But if she says exactly 89, then her opponent must say at least 90 and no more than 99. In all cases of her opponents move being between 90 and 99 she wins. Thus, saying 89 is a sufficient condition for guaranteeing herself the victory.

By the same logic, a player could say 89 in move $n-1$ if she said 78 in her move $n-2$. And so on. Going back to the beginning one can reconstruct a winning strategy that recommends picking 1, 12, 23, 34, 45, 56, 67, 78, 89, 100 regardless of the other player's choices. Thus, Player 1—who can start by saying "one"—can claim the victory.

Our solution obtained above clearly uses some form of backward induction. However, a more careful examination of the reasoning that we applied reveals that in every step that takes us backwards, we remove more than just one decision node with corresponding branches. For instance, in the last step, after Player 1 says 89, Player 2 could say any number from 90 to 99 before Player 1 says 100. All such numbers represent one or more subgames. Thus, in our solution we remove several subgames and several decision nodes at the same time. More specifically, we remove an entire complex subgame that starts with the decision node following Player 1 saying 89.

BI in the version described above appears to be incredibly helpful for solving a variety of relatively simple board games such as Tic Tac Toe or the following game of Nim:

> There are two players and two sets of beans, say, A and B. Both sets include exactly two beans. First, Player 1 takes away any number of beans from a set of his choice. Next, Player 2 takes away any number of beans from any set and so on. The player who is forced to take away the last remaining bean (or beans) is the loser.

The reader should be able to apply backward induction and identify the player who can always win by playing optimally. Then, the reader may increase the difficulty by adding beans, sets and players...

Kaminski (2009) implies that the above described procedure or removing multiple decision nodes with the branches originating in them is a legitimate extension of BI. We can substitute any disjoint subset of subgames with the SPE payoffs in those subgames and obtain all SPEs in the original game in the same way as it happens with the original BI. This property becomes critical for games with infinite actions or of infinite length.

Ultimatum game: Two players want to divide between themselves the amount of a good equal to 1. Player 1 chooses a number $x \in [0,1]$ that represents the amount Player 1 wants to keep. Next, Player 2 makes a choice. She may accept the proposed division and receive $1-x$ while Player 1 receives x, or she may reject it, in which case both players receive zero. Thus, Player 2 has veto power over the proposed allocation but applying the veto annihilates the reward for both players.

The BI reasoning stipulates that Player 2 accepts any positive amount offered to her and that she is indifferent between accepting and rejecting zero. Thus, in SPE Player 1 can propose only one, i.e., everything for himself. The unique SPE in the Ultimatum game is when Player 1 proposes one and Player 2 accepts any proposed amount, including zero.

The Ultimatum game provides a simple illustration of the usefulness of BI outside of the realm of finite games. However, as discussed later, its predictive power is questionable. A better example of how the strategic intuition motivating BI allows gaining insights into an important political process is provided by the next example.

Agenda-setter game (Romer and Rosenthal, 1978): Two players, the Agenda setter A and the Legislator L, have Euclidean preferences in the issue space $[0,3]$ and the ideal points $a=0$ and $l=2$, respectively. Euclidean preferences define the payoffs as negatives of distances between a player's ideal point and the point in the issue space resulting from the strategy profile. For Player A, the payoff associated with the final outcome w is simply $-|w|$ while for Player L, the payoff associated with w is $-|w-2|$.

The status quo is $q=3$. First, A proposes a policy $x \in [0,3)$. Next, L exercises his veto power and chooses the final law from $\{x,q\}$, i.e., either rejects (R) or agrees that the new policy be implemented (P).

To solve the game by BI, let's consider L's problem. Since the distance between l and q is 1, L prefers any new policy that is closer to l than 1; is

indifferent between 1 and q; and prefers q to any policy that is farther from l than 1. Thus, the best action responding to $x \in [0, 1)$ is R; the best answer to $x \in (1, 3)$ is P; and for 1 both R and P are best answers. Now, let's consider A's problem. Anticipating L's response, A does not try to impose on L his ideal point, 0. He offers a policy that is closer to L's ideal point. The policy that is closest to A's ideal point and acceptable for L is 1. Thus, the unique outcome selected by BI is 1.

While there is exactly one subgame perfect equilibrium, the Agenda-setter game has a huge number of Nash equilibria based on incredible threats. For instance, in one such equilibrium, L may threaten to veto everything except for his ideal point $l = 2$ and Agenda setter may propose 2. Of course, the veto threat is incredible since if A proposes, say, 1.5, then L has no incentive to carry out the veto and receive the status quo 3 instead of the better 1.5.

The Agenda-setter game explains a vast variety of strategic interactions taking place between the two chambers of parliament, parliaments and supreme courts or presidents endowed with veto power all over the world. It can be used to explain the practically universal introduction of lustration laws by ex-communist parties in transitional democracies of Central Europe (Kaminski and Nalepa, 2014). In all such cases, the first-moving player anticipates the opponent's response and preemptively chooses not his best option but rather the best option that will survive the potential veto of the opponent. While the calibration of issue spaces is a perennial problem, the universally observed phenomenon is that policy proposals take into account the preferences of the relevant veto players, and incorporate them into the offer.

Repeated Prisoner's Dilemma (PD): Assuming that the reader is familiar with the well-known anecdote behind the PD (due to Tucker, 1950), I will only briefly recall the point. The PD tells the story of two partners in crime who were caught by the police. While the Prosecutor cannot prove their crime without the cooperation of at least one of them, she can punish them with a lesser charge if they stay silent (cooperate). She can also offer a plea bargain to the one who will talk (defect).

The only Nash equilibrium in the PD is when both players play D. In fact, this unique NE has the very appealing property that the player strategies are dominant, i.e., they always give a player more than any other strategy. But what happens when we repeat the PD a finite number of times?

The original BI algorithm is not applicable to the repeated PD since the game involves some, arguably small, amount of imperfect information but

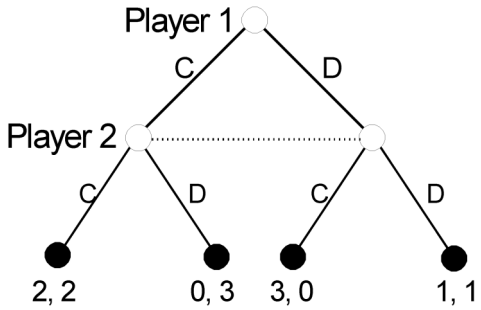

Figure 4. Prisoner's Dilemma

the generalized algorithm incorporates this case. We can substitute every repeated iteration of PD at the end of the game with the SPE equilibrium in this iteration, i.e., DD. Thus, the unique SPE (and also NE) in the finitely repeated PD is Always D for both players.

Criticism

It is sometimes argued that in the Ultimatum game that Player 1 very rarely makes zero offers, and that Player 2 sometimes rejects even positive amounts offered to them. However, the phenomenon may be explained not by the failure of BI to predict correctly the outcome, but rather by an incorrect specification of player payoffs in the experimental setting. The actual player payoffs may represent not only the amount that the players receive but also their feelings of fairness, the pressure associated with the experimenter, or the presence of the other player. When monetary rewards are small, as is the case in experiments, such feelings may seriously disturb the payoffs.

Another criticism is related to a popular BI anecdote, often called the Hangman Paradox or Unexpected Hanging, that tells the following story. The prison warden announces to a game theorist on death row: "Today is Sunday. You will be hanged during the next week. You will be very surprised when the day of your execution comes." The game theorist thinks for a while and responds with relief: "In such a case, I won't be hanged at all!"

What was his reasoning?

It is easy to reconstruct the poor game theorist's way of thinking. He cannot be hanged on next Sunday, the last day his execution is possible, since in such a case he wouldn't be surprised at all. Thus, since next Sunday is impossible, the last possible day of execution is Saturday. But it

is also impossible to hold the execution on Saturday since once it comes, it would be the last possible day for the execution because Sunday had already been eliminated. But this means that our game theorist would certainly expect execution on Saturday and wouldn't be surprised. By going backward, we can eliminate every single day in the next week. The execution is not possible.

How did the story end? The game theorist was hanged on Wednesday and he was very surprised!

The discussion of the Paradox is quite substantial (see e.g. Gardner, 1963) but the flaw in the backward reasoning described above is at the same time quite simple. The Paradox does not refer to any specific game (the reader may try to construct a game that would correspond to the story). Instead, it uses some concepts that do not even have formal equivalents in extensive form games, such as "being surprised." Thus, while there is some superficial similarity to BI, the backward reasoning is not BI. There is some reduction applied and the inference goes backward, but this is about where the similarity between the two types of reasoning ends.

There are more serious problems with BI than its alleged failure to predict the outcome of the Ultimatum game or explain the Hangman Paradox. One may ask a simple question: What would happen with games like chess if all potential players were using backward induction in a consistent fashion? The answer is now truly troublesome: nobody would be willing to play a game. Perfect "backward inductionists" would always reach the same outcome. We haven't learned so far whether it would be the victory of whites or blacks, or possibly a tie. What we do know is that in the world of perfect BI this determinate outcome would be well-known and there would be no thrill from playing the game the same way as there is no thrill from playing Pick-an-Integer game when you learn the winning strategy. Similarly, all board or computer games of perfect information would instantly lose their appeal.

Obviously, people play chess, checkers, Chinese checkers, Go, and other games and reach various outcomes, which implies that at least some of them are not perfect backward inductionists!

In fact, there are quite simple games in which real-world players practically always deviate from the principles of BI. The deviations can be attributed to the complexity of the decision environment, the presence of a small amount of incomplete information, or making assumptions about other players' reasoning not being perfectly compatible with BI. In games such as the chain-store paradox (Selten, 1978; Rosenthal, 1981), competing firms may engage in behavior that contradicts the predictions drawn from BI. Consider the following game of Centipede (see Figure 5 Centipede)

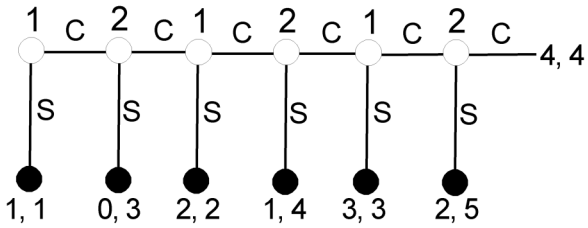

Figure 5. The game of Centipede with three rounds

Player 1 begins. She can continue or she can stop. When she stops, both players receive the payoff of one. If she continues, she increases the payoff of Player 2 by two and decreases her own payoff by one, assuming that Player 2 would immediately stop. Then Player 2 faces a similar choice, and either receives immediate reward or increases the reward of Player 1 by 2 at his own cost of 1. The game continues until some player stops.

Let's solve the game with backward induction: Player 2 in the last round receives 5 for stopping and 4 for continuing. Thus, when Player 2 stops, Player 1 anticipates such an action, and expects the payoff of 2 for continuing and 3 for stopping, and also stops. One can go all the way back to the first decision made by Player 1 and conclude that both players have always an incentive to stop. The unique BI solution to the Centipede is that Player 1 stops immediately and that both players would stop at any moment of the game.

However, one can notice that if both players continue for just one round, they will always get at least as much as in the unique BI equilibrium (in the worst-case scenario, Player 1's payoff drops to 1 while Player 2's payoff always remains strictly higher than 1). If they continue for at least two rounds, they always get strictly more. In fact, in experiments players typically recognize the rewards coming from a longer play and practically always continue until nearly the end of the game. (When demonstrating Centipede to his students, the author of the present article started with ten-rounds-long Centipedes and payoffs in real dollars. He quickly learned that the three-round version is sufficient to illustrate the problem and, needless to say, much less expensive!)

The epistemic literature on BI tries to explain the outcomes observed in Centipede by placing emphasis on player expectations. If Player 2 gets the chance to make his first move, he understands that Player 1 deviated from the reasoning prescribed by BI and didn't stop. Thus, he may be willing to assume a substantial probability of another future deviation, and continue

herself. But Player 1 can predict this reasoning of Player 2 and continue in the first round precisely because of such expectations.

Centipede is a game closely related to the Prisoner's Dilemma (see Figure 4 in the previous section). The application of the generalized BI procedure implies that players playing according to the principles of BI should always defect.

In reality, the deviations from the predictions made by BI were discovered in the very first game-theoretic experiment ever conducted and that defined the Prisoner's Dilemma! Flood and Dresher (Flood, 1952) introduced the PD game in an experiment in which they asked two players who were "well-familiar with two-person zero-sum game theory" (Flood, 1952, p. 17) to play one hundred repetitions of the PD. They noted that, curiously, players tended to get involved in cycles of cooperation and defection. The deviation from total defection was substantial. Needless to say, deviations from total defection persist if the experiment is repeated even among players who are not so "well-familiar with [...] game theory."

Another problem of BI is closely connected to the insufficient power of subgame perfection to eliminate unreasonable Nash equilibria. In the game shown below, there are three Nash equilibria (see Figure 6):

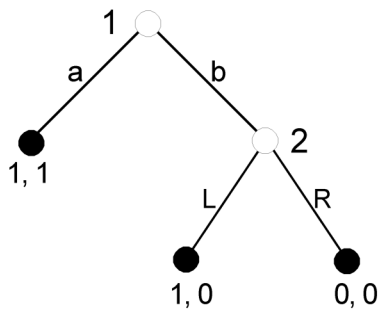

Figure 6. A game with BI solution that includes a weakly dominated strategy

Player 2 is indifferent between the two available actions L and R. The version of backward induction that is suitable for such cases would admit all three Nash equilibria: When 2 selects R, 1 selects a (equilibrium (a,R)); When 2 selects L, 1 selects either a or b (equilibria (a,L) and (b,L)). However, the third equilibrium (b,L) does not seem reasonable since the strategy b of player 1 is weakly dominated by his strategy a. One can hardly justify the use of this particular strategy by Player 1. Even if b performs against L as well as a, why would he take the risk that Player 2 would choose R

and bring him a lower payoff? Player 2 is indifferent between L and R and Player 1 has no reason to believe that one will be chosen over another.

The unreasonable looking strategy profile (b,L) that is admitted by BI can be eliminated by solution concepts such as perfect equilibrium that are more restrictive than SPE (Selten, 1975).

The predictions of BI are clearly more in tune with empirically observed or experimentally induced behavior for short games and a small number of players. Among the explanations of such deviations one finds modifications of the idea of rationality (Bonnano, 1991) expectations of the opponent's irrationality (Basu, 1988, 1990; Selten 1978), resorting to belief or the small amount of incomplete information that appears automatically in players' minds when the game is complicated enough (Petit and Sugden, 1989) or using a decision-theoretic approach (Rosenthal, 1981). None of those explanations was universally accepted. What we know for sure is what we see in experiments: adding complexity and length to some games may lead players to deviate from BI-consistent behavior.

The author is grateful to Barbara Kataneksza for comments and to the Center for the Study of Democracy for support.

REFERENCES

Basu, K. (1988). Strategic Irrationality in Extensive Games. *Mathematical Social Sciences* 15: 247–260.

— (1990). On the Non-Existence of a Rationality Definition for Extensive Games. *International Journal of Game Theory* 9: 33–44.

Binmore, K. (1987). Modeling Rational Players, Part 1. *Economics and Philosophy* 3: 179–214.

— Modeling Rational Players, Part 2. (1988). *Economics and Philosophy* 4: 9–55.

Bonanno, G. (1991). The Logic of Rational Play in Games of Perfect Information. *Economics and Philosophy* 7: 31–65.

Cournot, A. (1838). *Recherches sur les principes mathématiques de la théorie des richesses.* Paris: Chez L. Hachette.

Flood, M. (1952). Some experimental games. *RAND memorandum.* http://www.rand.org/content/dam/rand/pubs/research_memoranda/2008/RM789-1.pdf.

Fudenberg, D. and J. Tirole. (1991). *Game Theory.* Cambridge, MA: The MIT Press.

Gardner, M. (1963). The Paradox of the Unexpected Hanging. *Scientific American* 3.

Kaminski, M.M. (2009). Backward Induction and Subgame Perfection. IMBS working paper, University of California, Irvine, 09–01.

Kaminski, M.M. and Nalepa M.A. (2014). A Model of Strategic Preemption: Why do Post-Communists Hurt Themselves? *Decisions* 21: 31–65.

Kuhn, H.W. (1953). Extensive Games and the Problem of Information. In H.W. Kuhn, and A.W. Tucker (Eds.), *Contributions to the Theory of Games 1* (pp. 193–216). Princeton: Princeton University Press.

Myerson, R. (1991). *Game Theory. Analysis of Conflict.* Cambridge, MA: Harvard University Press.

Petit, P. and Sugden, R. (1989). The Backward Induction Paradox. *Economics and Philosophy* 14: 95–125.

Romer, T. and Rosenthal, H. (1978) Political Resource Allocation, Controlled Agendas, and the Status Quo. *Public Choice* 33: 27–44.

Rosenthal, R.W. (1981). Games of Perfect Information, Predatory Pricing, and the Chain-Store Paradox, *Journal of Economic Theory* 25: 92–100.

Schelling, T.C. (1960). *The Strategy of Conflict.* Cambridge, Ma: Harvard University Press.

Schwalbe, U. and Walker, P. (2001). Zermelo and the Early History of Game Theory. *Games and Economic Behavior* 34: 123–137.

Selten, R. (1965). Spieltheoretische Behandlung Eines Oligopolmodels Mit Nachfrageträgheit. *Zeitschrift für die gesampte Staatswissenschaft* (pp. 667–689). Tübingen: Mohr Siebeck GmbH & Co. KG.

Selten, R. (1975). Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games. *International Journal of Game Theory* 4(1): 25–55.

Selten, R. (1978). The Chain Store Paradox. *Theory and Decision.* 9: 127–159.

Stackelberg, H.F.V. (1934). *Marktform Und Gleichgewicht (Market Structure and Equilibrium).* Vienna: J. Springer.

Tucker, A. W. (1950). A two-person dilemma. Mimeographed paper, Stanford University. Published in 1980 as: On jargon: The prisoner's dilemma. *UMAP Journal* 1:101.

von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior.* Princeton: Princeton University Press.

Zermelo, E. (1913). *Über Eine Anwendung Der Mengenlehre Auf Die Theorie Des Schachspiels.* Cambridge: Cambridge University Press, 501–504.