Backward Induction Lecture Notes

Note: These provide a lot of lecture material; I could have lectured for four hours and probably still had a bit left over.

Before-lecture preparation: Early in the semester, promise the students that you will bring them surprise cookies: they won't know what day the cookies are coming until the cookies actually arrive. A few days before the backward induction lecture, bring the surprise cookies. Ask them if they were surprised (they should say "yes"). Ask them if they remembered the original announcement (a fair number of them should say "yes"). Argue that they couldn't have known the announcement would be true. Since they know the cookies will be a surprise, they know the cookies can't come on the last day; since they know the cookies can't come on the last day and the cookies will be a surprise, they know the cookies can't come on the next-to-last day, etc.

In the lecture: get two volunteers to play the repeated prisoner's dilemmma game. At each round, the players do a game with this matrix:

		Player 1	
		Cooperate	Defect
Player 2	Cooperate	3/3	5/0
	Defect	0/5	2/2

Do students defect on the last several rounds? Discuss why defection does/doesn't happen.

Go over Carroll's backward induction argument in the iterated prisoner's dilemma, from these premises:

Condition (G)

(a) Players 1 and 2 are playing a sequence of exactly *n* prisoner's dilemmas
(b) Both players are rational throughout the sequence
(c) For all *k* players correctly believe just before the *k*th prisoner's dilemma what moves were played on the 1^{st} through k-1^{st} prisoner's dilemma.

(B1) Both players know (G)
(B2) Both players know (B1)
…
Etc.

Explain relevant sense of "rational" as causal decision theoretic rationality: neither player will make a sacrifice unless it will *cause* her to gain something in the future.

Discuss difference between reading premises as true at initial stage (plausible, but argument is invalid) vs. true throughout (argument is valid, but premises are implausible).

Search for a way of weakening the premises to get plausibility + validity.

- If no suggestions, propose Broome/Rabinowicz solution, as follows

0) At each round in the game that has been reached without an irrational move, the player will act rationally.
1) At each round that has been reached without an irrational move, both players believe (0).
2) At each round that has been reached without an irrational move, both players believe (0).
…
Etc.

(1) and all succeeding premises follow from these premises:

A) At the beginning of the game, players have no false beliefs.
B) At no round do players acquire false beliefs.
C) Both players retain their beliefs for as long as they are consistent with their acquired beliefs.
D) At the beginning of the game, there is common belief in (0) and (A)-(C)

What is a Nash equilibrium, and why doesn't Carroll care about them?

- Explain Nash Equilibria using matrices for Chicken and Prisoner's dilemma.
  - Chicken has two equilibria:

		Player 1	
		Swerve	Stick
Player 2	Swerve	2/2	**10/0**
	Stick	**0/10**	-10/-10

- A game with a Nash equilibrium that isn't very worthwhile:

50/70	60/60	70/50	4/1
60/60	40/40	60/60	3/1
70/50	60/60	50/70	8/1
1/5	1/6	1/7	**2/2**

Large group activity: adapt Carroll's backward induction argument to the surprise cookies case.

- Are the analogous premises more/less plausible in the surprise cookies case than in the prisoner's dilemma case?
- Can we avoid the backward induction if we discard positive introspection?
  - Positive introspection = "P is believed → It is believed that P is believed"
- A version of the sorites argument that uses positive introspection:
  - If I know that F(n) [e.g., that the tree outside is not *n* inches tall; that a man with 0 hairs is not hairy], then it's not the case that F(n+1).
  - I know that F(0).
  - Therefore, I know that F(n-1).
- Discussion: can we solve the problem of vagueness by discarding positive introspection?

Why isn’t the teacher’s announcement paradoxical if the students overhear it, instead of being told?

- Segue to the suggestion that successful assertion makes for common knowledge of asserted content.
    - Clumsy waiter example: waiter spills a glass of water and declares, “It’s my fault”; this can be useful even if both parties know that the spill is the waiter’s fault.
- Can there be common knowledge of the premises in the surprise exam/cookies case?
- Common knowledge game: Pick two volunteers. Volunteers must choose the same number from the set {1, 2, 3, 77} without communicating about their choices. If they coordinate, each wins a brownie; otherwise, there is no prize.
- Next, have students choose from {1, 2, 3, 4}. If they coordinate on 1,2, or 4, each wins a brownie; if they coordinate on 3, each wins half a brownie; otherwise, there is no prize.
- Discussion: how did students coordinate? Was it easier to coordinate in one of the games than in the other? If so, why?