DS 574 Algorithmic Mechanism Design	Lecture #1
Fall 2026	Prof. Kira Goldner

Before We Start: Write Down a Number

Exercise.

a. Write down your name and a whole number between 0 and 100, inclusive.

b. The winner is whoever is closest to $\frac{2}{3}$ of the *average* of every number written down in this room.

c. Enter it on the Google Form via QR code. Don't discuss it with your neighbor.

Eighty Minutes in Gijón

On June 25, 1982, West Germany played Austria in Gijón in the last match of the World Cup group stage. In 1982, during the group stage, there was a round robin (everyone played each other) where wins were worth 2 points, draws worth 1 point, and losses worth 0.

Date	**Match**	**Result**
16 June	Algeria–West Germany	2–1
17 June	Austria–Chile	1–0
20 June	West Germany–Chile	4–1
21 June	Austria–Algeria	2–0
24 June	Algeria–Chile	3–2

Table 1: Group 2 results before the final match.

Team	**P**	**W**	**D**	**L**	**GF**	**GA**	**GD**	**Pts**
Austria	2	2	0	0	3	0	+3	4
Algeria	3	2	0	1	5	5	0	4
West Germany	2	1	0	1	5	3	+2	2
Chile	3	0	0	3	3	8	−5	0

Table 2: Group 2 at kickoff. Algeria and Chile have completed their matches.

Two teams advance determined by those with the most points. If teams tie on points, the next differentiator was goal differential (goals scored minus goals let in). Only West Germany and Austria are left and can do the math, so what should they do?

Result of West Germany–Austria	**Who advances**
West Germany wins by 3 or more	West Germany & Algeria advance
West Germany wins by 1 or 2	**West Germany and Austria both advance**
Draw	Austria & Algeria advance
Austria wins	Austria & Algeria advance

There is exactly one outcome that saves both teams, and both teams know it before the whistle.

So what do they do? West Germany scored in the 10th minute, 1–0. And then, for eighty minutes, the two teams passed the ball sideways. Both European teams advanced. Algeria—who had beaten West Germany and Chile—went home.

Let's formalize what "utility" means for each team under each result. Write x for the result of the match and X for the four rows above. Give each team a reward

$$u_i(x) = \begin{cases} 1 & \text{if team } i \text{ advances under result } x, \\ 0 & \text{otherwise,} \end{cases} \qquad i \in \{\mathrm{WG}, \mathrm{AUT}\}.$$

Then a "best result" for a team is one that maximizes its utility: a best result for team i is a result in $\operatorname{argmax}_{x\in X} u_i(x)$. Write $x^\star$ for "West Germany wins by 1 or 2."

Claim 1. $x^\star$ is a best result for West Germany *and* a best result for Austria, and it is the only result with that property. That is, $\operatorname{argmax}_{x\in X} u_{\mathrm{WG}}(x) \cap \operatorname{argmax}_{x\in X} u_{\mathrm{AUT}}(x) = \{x^\star\}$.

Proof. Both rewards take values in $\{0, 1\}$, so 1 is the largest value either can take. At $x^\star$ both equal 1, so $x^\star$ maximizes both. Now take any other $x \in X$. Since $u_{\mathrm{WG}}(x) \cdot u_{\mathrm{AUT}}(x) = 0$, at least one of the two factors is 0, and that team strictly prefers $x^\star$ to x. So x is not in the intersection. □

Note that this means that the outcome is *collusive* (there is a shared best result) as opposed to competitive [Csató and Gyimesi, 2026].

> **Exercise.** You are FIFA, in July 1982, and this must not happen again. Write down *one* concrete change to the rules.

Two kinds of rule change. The first punishes: fine both teams; dock points; disqualify a team that "fails to compete." It is the natural instrument and it is the wrong one. What would the rule say? Something like "a team that does not try to win is punished." Now: how would you know? You have a 1–0 scoreline and eighty minutes of sideways passing—which is also what a team does when it is protecting a lead, which is legal, normal, and often correct. *Trying* is not something a referee can observe. Outcomes are observable; effort is not, and a rule written in terms of effort mostly deters the honest and the clumsy.

The second kind changes the game, so that the shared favorite outcome is no longer something the two teams can *steer toward*. From the 1986 World Cup onward, the final pair of matches in every group kicks off simultaneously. This is exactly what FIFA did: they did not add a rule about behavior, did not add a penalty, and did not ask anyone to be virtuous. They deleted a piece of information.

But the original incentive is still there. Simultaneous kickoff does not change anybody's preferences—two teams that would both like to advance would still both like to advance. What it destroys is not the incentive but the *guarantee*.

One postscript, because fixes have side effects. Csató and Gyimesi [2026] look forward at the 48-team format used in 2026 and find that last-round matches where neither side has much riding on the result roughly *double*, from under 20% to almost 40%—and that preferring head-to-head over goal difference as the first tiebreaker, the intuitive anti-collusion choice and the one FIFA made, *increased* unimportant matches by about 5.6 percentage points.

Two-Thirds of the Average

Back to our numbers. The average, then two-thirds of it, then the winner.

What did you actually do to pick your number? Three kinds of answers:

- "I picked the average of 0 to 100, so 50."
- "I figured people would say 50, so I said 2/3 that, which is 33."
- "I figured people would figure that, so I said 2/3 that, which is 22."

This is three *depths* of the same move—reasoning about what everybody else is going to reason.

The model. Each of the n people in the room picks a number $b_i \in [0, 100]$. Call b_i player i's *strategy*, and call the vector $b = (b_1, \ldots, b_n)$ of everyone's choices a *strategy profile*. The target is $t(b) = \frac{2}{3} \cdot \frac{1}{n} \sum_{j=1}^{n} b_j$. Player i's *payoff* is $-|b_i - t(b)|$: closer is better. Note: your own number is part of the average.

Definition 1. Strategy b_i is *strictly dominated* by strategy b_i' if, no matter what everybody else does, i's payoff from b_i' is strictly larger than i's payoff from b_i.

That is, a strictly dominated strategy is one you can throw away without knowing anything at all about the other players—no beliefs, no psychology, no trust.

Theorem 1. *Suppose every player's number lies in* $[0, M]$ *for some* $M > 0$*. Then for player* i*, every strategy* $b_i > \frac{2}{3}M$ *is strictly dominated by the strategy* $\frac{2}{3}M$*.*

Proof. Call avg the average. Note that since every individual $b_i \leq M$, then $avg \leq M$, so $2/3avg \leq 2/3M$. Hence there is no reason to ever bid over $2/3M$; this is strict dominance. $\square$

Corollary 2. *Applying Theorem 1 with* $M_0 = 100$*, no rational player picks above* $M_1 = \frac{2}{3} \cdot 100 = 66.\overline{6}$*. If nobody picks above* M_1*, the theorem applies again with* $M = M_1$ *and nobody picks above* $M_2 = \frac{2}{3}M_1 = 44.4$*; then* $M_3 = 29.6$*,* $M_4 = 19.8$*, and in general* $M_k = 100\left(\frac{2}{3}\right)^k \to 0$*. The only number surviving every round of this elimination is* 0*.*

A couple of people usually write 0, and they usually lose. Class averages in this game usually land somewhere in the 20s to 40s, and the winning number with them. Playing 0 is correct only if you believe everybody else will do the full elimination too—and Corollary 2 tells you what the *fully rational* room does, not what *this* room does.

> **Exercise.** Let's play *again*, right now, same rules, having just seen the tally. Write a new number and pass it forward.

The average falls, by a lot, but not to zero. The take away: a mechanism has to work on the players you actually have, not on the ones your theorem assumes. Note: Alekseenko et al. [2025] ran this exact experiment with LLMs as the players—they find most models play *lower* than humans do, and still fail to play dominant strategies in a two-player version.

Split or Steal

Golden Balls ran on ITV from 2007 to 2009. In the last sixty seconds of an episode, two contestants have built a jackpot of J pounds. Each secretly picks a ball marked *Split* or *Steal*. Then if: both Split, and they split the jackpot, $J/2$ each; one Splits and one Steals, and the Stealer takes all of J while the Splitter gets nothing; both Steal, and both get nothing.

> **Exercise.** In pairs, competing for a jackpot, you may talk to each other for ninety seconds, and then you choose simultaneously.

Now let's write it down. Rows are your choice, columns are theirs, and each cell is (your payoff, their payoff):

	They Split	**They Steal**
You Split	$(J/2,\ J/2)$	$(0,\ J)$
You Steal	$(J,\ 0)$	$(0,\ 0)$

Definition 2. Strategy b_i is *weakly dominated* by b_i' if, no matter what everybody else does, i's payoff from b_i' is at least as large as from b_i, and for at least one choice by the others it is strictly larger.

Claim 2. In *Golden Balls* with jackpot $J > 0$, Split is weakly dominated by Steal. It is not strictly dominated.

Proof. Read the column marked "They Split": Steal pays you J and Split pays you $J/2$, and $J > J/2$ since $J > 0$, so Steal is strictly better. Read the column marked "They Steal": Steal pays 0 and Split pays 0, so the two are exactly tied. Steal is therefore never worse and sometimes strictly better, which is weak dominance; and because of the tie in the second column it is not strict dominance. □

Why do we care about the difference between weak and strict here? Because the tie in that second column is what the next example exploits.

The Nick Corrigan / Ibrahim Hussein episode had a jackpot of £13,600. In the talking phase, Nick announced, up front and flatly, that he was going to choose Steal, and then told Ibrahim: choose Split, and I'll give you half of it afterwards.

Work out Ibrahim's problem. Suppose Ibrahim believes Nick. Let $q \in [0, 1]$ be the chance that Nick actually hands over half afterwards—an unenforceable promise. Then Ibrahim's payoffs are

$$\text{Split:} \quad 0 + q \cdot \frac{J}{2}, \qquad\qquad \text{Steal:} \quad 0,$$

and Split is strictly better if and only if $q > 0$.

By committing to Steal before the choice, Nick converted a symmetric simultaneous game into a take-it-or-leave-it offer where Ibrahim's own arithmetic pointed at Split.

Commitment. Removing one of your own options can make you better off. The auctioneer who publishes a reserve price and cannot be talked down, the central bank that ties its own hands, the platform that publishes its ranking algorithm.

Nick played Split. Ibrahim played Split. They took £6,800 each, which is exactly what the promise said they'd get. Asked afterwards whether he had ever considered splitting, Ibrahim said: "No. Not at all. Not at all. I was always gonna steal." The commitment didn't only get Nick a good outcome. It got Ibrahim a good outcome, against Ibrahim's own plan.

What We Observed Today

Definition 3. A *mechanism* is a rule x that maps the players' actions $b = (b_1, \ldots, b_n)$ to an outcome $x(b)$, together with the rewards each player receives from that outcome.

Mechanisms include: tournament formats, points formulas, auctions, school-assignment systems, ballots, promotion policies, recommendation algorithms, etc. Designing one means choosing x knowing that the players will choose b in response to x.

So the organizing question of this course is: *when can we design a rule so that the behavior we want is the behavior the players choose?* Sometimes we can, exactly and provably. Sometimes we can prove that we cannot, which is its own kind of useful. And sometimes—Gijón—we can only see afterwards which one we were in.

And another observation of today is: **if losing can help you, someone will lose on purpose, and blaming them is a category error. The rule did that.** "The players should have tried harder" is not a fix; it is a way of not writing one.

Note: Everything we did today had visible stakes: the rule is published, the actions are on television, and the reward is a trophy. Many things we'll study will not be as full information.

Acknowledgements

This lecture was developed in part using materials by Tim Roughgarden, and in particular, his book “Twenty Lectures on Algorithmic Game Theory” [Roughgarden, 2016]. The last-round match classification and the 48-team World Cup figures are from Csató and Gyimesi [2026].

References

Iaroslav Alekseenko, Dmitry Dagaev, Sofia Paklina, and Petr Parshakov. Strategizing with AI: Insights from a beauty contest experiment, 2025. arXiv preprint arXiv:2502.03158.

László Csató and András Gyimesi. Match classification in the last round of four-team round-robin tournaments, 2026. arXiv preprint arXiv:2605.25610.

Tim Roughgarden. *Twenty Lectures on Algorithmic Game Theory*. Cambridge University Press, 2016.