CS 405/6001: Game Theory and Algorithmic Mechanism Design **Jul–Nov 2026**

Lecture 1: Why Game Theory, and the Game of Chess

Lecturer: Swaprava Nath *Scribe(s): Claude Scribe*

Disclaimer: *These notes aggregate content from several texts and have not been subjected to the usual scrutiny deserved by formal publications. If you find errors, please bring to the notice of the Instructor.*

1.1 Two Ways to Study a System

Engineering courses usually split a subject into two dual problems. In circuit theory, *analysis* takes a given circuit and works out its behavior; *synthesis* starts from a desired behavior and designs a circuit that produces it. The same duality organizes this course, except the "circuit" is a group of self-interested people.

Definition 1.1 ***Game theory*** *takes a given set of players, their available actions, and their preferences over outcomes, and predicts what a* rational *and* intelligent *player will do.*

Definition 1.2 ***Mechanism design*** *starts from a desired social outcome (fairness, efficiency, revenue, truthfulness, . . .) and designs the rules of a game so that self-interested, rational play produces that outcome. It is sometimes called* inverse game theory*.*

Mechanism design as a distinct field is usually traced to Hurwicz's work on resource allocation mechanisms [Hur73]; Hurwicz, together with Maskin and Myerson, received the 2007 Nobel Memorial Prize in Economic Sciences for laying its foundations. We will use both directions throughout the course: predicting outcomes of a given game (Weeks 1–7, roughly) and designing games with guaranteed properties (auctions, voting, matching – the second half of the course).

1.1.1 A First Game

Consider two AI labs, Alpha and Beta, racing to release a new model. Each can either Test Thoroughly (slow, safe) or Ship Fast (skip long safety evaluations). Suppose the payoffs (utilities) are as in Table 1.1.

		Beta	
		Test Thoroughly	Ship Fast
Alpha	Test Thoroughly	$5, 5$	$0, 6$
	Ship Fast	$6, 0$	$1, 1$

Table 1.1: The AI Lab Dilemma.

This is our first example of a *game*: a set of players $N = \{\text{Alpha}, \text{Beta}\}$, a set of actions per player, and a payoff to every player for every combination of actions chosen. We will formalize this as a *strategic-form game* in the Lecture 2 notes, and revisit exactly this table under its textbook name, the *Prisoner's Dilemma*.

1.1.2 Rationality, Intelligence

Definition 1.3 *A player is **rational** if she picks an action to maximize her own payoff (utility), given her beliefs about what the other players will do.*

Definition 1.4 *A player is **intelligent** if she knows the rules of the game perfectly and can compute a best action, knowing that the other players are also rational and intelligent.*

Remark 1.5 *"Maximizing utility" is not an arbitrary modeling choice. Under the von Neumann–Morgenstern axioms on a player's preferences over lotteries (completeness, transitivity, continuity, independence), preferences can be represented by a real-valued utility function, unique up to positive affine transformation, whose* expectation *the player maximizes [MSZ13ch2]. We will not need the axioms until we introduce mixed strategies in a few weeks, but it is worth knowing that "rationality = utility maximization" rests on a theorem, not a stipulation.*

Game theory's foundational assumption is that rationality and intelligence are not just true, but *commonly known* to be true. This turns out to be a surprisingly subtle idea, which we unpack next with a classical puzzle.

1.2 Common Knowledge

Definition 1.6 *A fact F is **common knowledge** among a group of players if: everyone knows F; everyone knows that everyone knows F; everyone knows that everyone knows that everyone knows F; and so on, ad infinitum.*

The mathematical study of common knowledge in games and economics was put on a rigorous footing by Aumann [Aum76], who showed, among other things, that two Bayesian agents with the same prior cannot "agree to disagree" once their posteriors are common knowledge. A comprehensive treatment of the logic of knowledge and common knowledge, with applications to distributed computing as well as game theory, is Fagin, Halpern, Moses and Vardi's textbook [FHMV95].

1.2.1 The Blue-Eyed Islanders

Puzzles of this flavor appear at least as early as Littlewood's collection of mathematical problems [Lit53]; the version below is the one used in class.

Setup. On an isolated island, with no mirrors and no discussion of eye color, live n people, each with either blue or black eyes. A visiting outsider, known to always tell the truth, announces to everyone simultaneously: "At least one of you has blue eyes." Any islander who deduces that his own eyes are blue must leave that same night. Everyone reasons perfectly and this is common knowledge.

Claim 1.7 *If there are exactly $k \geq 1$ blue-eyed islanders, all of them leave on night k, and no one leaves before then.*

Proof: By induction on k.

Base case ($k = 1$). The one blue-eyed islander sees only black eyes around him. Since the outsider is truthful, and no one else can be the blue-eyed person referred to, he concludes his own eyes must be blue, and leaves on night 1.

Inductive step. Suppose the claim holds for $k-1$. If there are exactly k blue-eyed islanders, each of them sees exactly $k-1$ other blue-eyed islanders. By the inductive hypothesis, if there were only $k-1$ blue-eyed people, they would all have left on night $k-1$. Every blue-eyed islander can run this argument, and, being intelligent, knows the others can too; when night $k-1$ passes and no one has left, each of the k blue-eyed islanders concludes that there must be a k-th blue-eyed person besides the $k-1$ he sees – namely, himself. All k leave on night k.

Nobody leaves *before* night k: an inductive argument identical to the above (omitted) shows that with k blue-eyed islanders, no contradiction with the outsider's statement arises until night $k-1$ has passed without a departure. ■

Remark 1.8 *The puzzle is often mistaken for a paradox, since every islander can already* see *who has blue eyes; nothing about the raw information changes when the outsider speaks. What changes is that the fact becomes* common *knowledge: before the announcement, each blue-eyed islander knew someone had blue eyes, but did not know that everyone else knew that he knew it, and so on. The outsider's public announcement is precisely what promotes shared private information to common knowledge, and it is only common knowledge (not merely widespread individual knowledge) that lets the induction above get off the ground.*

This is exactly the assumption our definition of game theory relies on: it is not enough that players are rational and intelligent; it must be common knowledge that they are, or a player cannot safely reason about what the others will do.

1.3 Mechanism Design, Warmed Up

Example 1.9 (Divide and Choose) *Two labmates, Riya and Karan, share a fixed budget (say, of cloud GPU credits) that must be split between them. Neither trusts the other's private valuation of "useful compute," and their advisor does not want to arbitrate. Consider the following procedure: Riya splits the budget into two shares, however she likes; Karan then picks whichever share he prefers, and Riya receives the other.*

Claim 1.10 *The divide-and-choose mechanism produces an outcome that is* ***envy-free****: neither player, judged by her own valuation, prefers the other's share to her own.*

Proof: Karan cannot envy Riya: he picked the share he valued at least as much as the other, by construction. Riya cannot envy Karan either: not knowing which share Karan will take, her best response is to divide the budget into two shares that she *herself* values equally – otherwise Karan might take the one she values more, at her expense. Having equalized the two shares by her own valuation, she is indifferent between them, and so does not envy whichever one Karan leaves her. ■

This two-player fair-division procedure goes back to Steinhaus's short note formalizing the "cake-cutting" problem [Ste48]; a full treatment of fair division, including n-player generalizations of divide-and-choose, is Brams and Taylor's book [BT96]. Notice the mechanism-design flavor: the advisor never needs to learn either student's private valuation. The rules of the game do the work.

1.4 The Game of Chess, Formally

We now build the vocabulary needed to state a nontrivial theorem about a familiar game. This section follows Maschler, Solan and Zamir's textbook [[]Ch. 1]MSZ13, which we abbreviate MSZ throughout these notes.

1.4.1 Setup

Chess is played by two players, White (W) and Black (B), each starting with 16 pieces; White moves first and the players alternate. A play ends in a win for W (B's king is captured), a win for B (W's king is captured), or a draw (stalemate, agreement, repetition, and several other rules we will not enumerate).

Definition 1.11 (MSZ, Def. 1.1) *A **game situation** is a finite sequence $(x_0, x_1, \ldots, x_K)$ of board positions such that x_0 is the opening position, and for each k, going from x_k to x_{k+1} is a single legal move of W if k is even, and of B if k is odd. Write H for the set of all game situations.*

A *board position* alone is not enough to reason about strategy, because several different move sequences can arrive at the same board; the full history matters for defining what a player's strategy prescribes at that point.

Definition 1.12 (MSZ, Def. 1.2) *A **strategy** for W is a function s_W that maps every game situation $(x_0, \ldots, x_k) \in H$ with k even to a board position x_{k+1} reachable from x_k by a single legal move of W. A strategy s_B for B is defined analogously for odd k.*

A pair of strategies (s_W, s_B) pins down a unique play of the entire game: $x_1 = s_W(x_0)$, $x_2 = s_B(x_0, x_1)$, and so on. The set of game situations forms a tree (the *game tree*), rooted at x_0, in which every leaf is a terminal situation (win, loss, or draw).

Definition 1.13 *A strategy s^*_W is a **winning strategy** for W if, for every strategy s_B of B, the play determined by (s^*_W, s_B) ends in a win for W. A strategy s'_W **guarantees at least a draw** for W if every play (s'_W, s_B) ends in a win for W or a draw. Winning and drawing strategies for B are defined symmetrically.*

1.4.2 An Early Theorem

Theorem 1.14 *In chess, exactly one of the following holds:*

(i) White has a winning strategy;

(ii) Black has a winning strategy;

(iii) each player has a strategy guaranteeing at least a draw.

MSZ attribute this result to von Neumann [[]Ch. 1]MSZ13. It is worth being precise about attribution here, since it is a common source of confusion.

Remark 1.15 (A historical aside) *The finite two-player perfect-information result of Theorem 1.14 is most often attributed in the game theory literature to Zermelo's 1913 address to the International Congress of Mathematicians [Zer13], and the result is frequently called* Zermelo's theorem*. Von Neumann's celebrated 1928 paper [vN28] instead proves the* minimax theorem *for two-player zero-sum games with simultaneous moves – a different (though related) result that we will need once we introduce mixed strategies. Schwalbe and Walker's historical study [SW01] carefully reconstructs what Zermelo's original paper actually proves (it is not quite Theorem 1.14 as stated either, and contains a gap later filled in by König and Kalmár) and untangles a genuinely confused citation history. The moral for us: know the statement and the proof cold; attributions in secondary sources, this one included, should be checked against primary sources before being repeated.*

The theorem is remarkable precisely because it rules out a fourth logical possibility: that *no* player can guarantee anything at all. It says nothing about *which* of (i)–(iii) holds for chess, nor what the guaranteeing strategy is – and, as of today, this remains unknown. We prove Theorem 1.14 by induction on the game tree, and give a second, purely logical proof, in Lecture 2.

1.5 Summary and Where We Are Headed

We introduced the analysis/synthesis duality that separates game theory from mechanism design; the rationality, intelligence, and common-knowledge assumptions underlying the entire course; a first fair-division mechanism (divide-and-choose); and the formal machinery (game situations, strategies) needed to state von Neumann's theorem about chess. In Lecture 2, we prove the theorem two different ways, meet the *strategic (normal) form* representation of a game that we will use for the next several weeks, and see several more games – some of which will look surprisingly familiar.

References

[Aum76] R. J. Aumann, "Agreeing to disagree," *The Annals of Statistics*, 4(6), 1976, pp. 1236–1239.

[BT96] S. J. Brams and A. D. Taylor, *Fair Division: From Cake-Cutting to Dispute Resolution*, Cambridge University Press, 1996.

[FHMV95] R. Fagin, J. Y. Halpern, Y. Moses and M. Y. Vardi, *Reasoning About Knowledge*, MIT Press, 1995.

[Hur73] L. Hurwicz, "The design of mechanisms for resource allocation," *American Economic Review*, 63(2), 1973, pp. 1–30.

[Lit53] J. E. Littlewood, *A Mathematician's Miscellany*, Methuen and Co., 1953.

[MSZ13] M. Maschler, E. Solan and S. Zamir, *Game Theory*, Chapter 1: "The Game of Chess," Cambridge University Press, 2013.

[MSZ13ch2] M. Maschler, E. Solan and S. Zamir, *Game Theory*, Chapter 2: "Utility Theory," Cambridge University Press, 2013.

[Ste48] H. Steinhaus, "The problem of fair division," *Econometrica*, 16(1), 1948, pp. 101–104.

[SW01] U. Schwalbe and P. Walker, "Zermelo and the early history of game theory," *Games and Economic Behavior*, 34(1), 2001, pp. 123–137.

[vN28] J. von Neumann, "Zur Theorie der Gesellschaftsspiele," *Mathematische Annalen*, 100(1), 1928, pp. 295–320.

[Zer13] E. Zermelo, "Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels," *Proceedings of the Fifth International Congress of Mathematicians*, Vol. 2, 1913, pp. 501–504.