A Teaser to Game Theory

Christian Hilbe, christian.hilbe@it-u.at
Version December 12, 2025

1 Motivation: About this class

Remark 1.1 (Background and motivation of this course)
This course is a short excursus within the PIC on Machine Learning and Natural Language Processing. It should expose students to different techniques to analyze strategic decision making. I did not *have to* give this course; I *wanted to*. With it, I want to achieve at least three goals:

1. Most immediately, it should provide a gentle introduction to game theory. Students will learn how to interpret real decision problems as a *game*. Then they will gain some intuition for how to find solutions to these decision problems (e.g., to compute the game's *equilibria*).
2. More generally, the course should serve as an example of mathematical modeling. Students should have in mind that the framework of game theory (the formalism, the solution concepts) are human constructs. They should understand why smart people at some point found it useful to come up with this formalism; they should be able to critically question the underlying assumptions and methods. Ideally, when taking this course, students should put themselves into the shoes of a smart social scientist, a hundred years ago (when game theory was not yet an established formalism). Ideally, they try to come up with some of the formalism and some of the solution concepts themselves. The formalism and the solution concepts can be tested empirically.
3. I hope the course will also just serve as brainfood. Students should enjoy to hear more about some scientific discipline that they might find stimulating, but that they would not hear about otherwise.

Remark 1.2 (Prerequisites)
A fully rigorous game theory course requires quite a bit of math (e.g., matrices, derivatives, fixed point theorems). However, for this lesson, it's not worth to develop the mathematical machinery. Instead, we entirely focus on intuitions. We introduce some key concepts without much mathematical formalism. However, I will suggest some reading material that interested students could use for further reading.

Remark 1.3 (What is game theory?)
Game theory is a scientific field that explores strategic decision-making in interactions with multiple decision-makers who jointly determine each others' outcomes. In particular, not all decision-making is covered by game theory; as an example, the following questions are not covered: *How to play roulette? Should I buy a piece of art at a given price? Should I do a PhD?* Instead, a strategic component is important. So, the following variations of the previous questions are in fact covered by game theory: *How*

to play poker? How much to bid for a piece of art within an auction? How to apply to a PhD program? In the following, we will ask two questions: (1) How can we formalize such strategic interactions in a mathematical model? (2) Given we have a model, what does it mean to *solve* a game?

2 Constructing a game-theoretical model

Example 2.1 (**An auction**)
Let's play the following game. Suppose there is a piece of art by a famous artist that you might want to buy. You know the value that this piece of art has to you (it is written on a Post-It). You don't know the value that other potential buyers assign to it. But you know how many other buyers there are, and you know that for each of them, their evaluation of the piece of art is randomly drawn from $\{$0 EUR, 1 EUR, 2 EUR, ..., 10 EUR$\}$. The rules of the auction are: Everyone secretly writes down the price they would be willing to pay (their *bid*). The potential buyer with the highest bid needs to pay the respective price, but gets the piece of art in return. (If there are multiple buyers with the same bid, it's randomly allocated). Question: *How much are you willing to bid?*

Remark 2.2 (**Elements of a game**)
When formalizing strategic interactions as a game, it is important to capture the following five elements:

- *Players.* Who are the relevant individuals that make (strategic) decisions?
- *Actions.* Between which actions can the players decide?
- *Order of moves.* Who gets to make a decision at which point?
- *Information.* What is it that players know when making their decisions?
- *Payoffs.* Depending on the players' decisions, what is the eventual outcome for each player? [e.g., measured in money, or happiness, or *utility*]

Example 2.3 (**Revisiting auctions**)
For our initial example of an auction, how would we define each of these five elements?

- *Players.* All the members of the classroom, $\mathcal{N} = \{1, 2, \ldots, n\}$.
- *Actions.* Each player needs to name a price they are willing to pay, $p_i \in \{0, 1, \ldots, 10\}$ for $i \in \mathcal{N}$.
- *Order of moves.* All players need to make their decision simultaneously.
- *Information.* Players know how their own private value of the good, $v_i \in \{0, 1, \ldots, 10\}$. They don't know the precise value of the other players, but they know it's uniformly distributed in $\{0, 1, \ldots, 10\}$.
- *Payoffs.* The player who gets the item has a payoff of $v_i - p_i$; all others have a payoff of zero.

Remark 2.4 (**Solving games**)
So far, we looked at how we can represent strategic interactions in more formal way. Next, we ask: What does it mean to solve such games (roughly: how should we predict what rational players would do). In this class, we will discuss two solution concepts: (1) (Iterated) elimination of dominated actions, and (2) Nash equilibria.

3 Solving (normal-form) games

Remark 3.1 (**Normal-form game**)
To get intuitions, we will look at the simplest possible case. We consider a scenario with two players only. Those players have complete knowledge of the possible actions and each other's possible payoffs. They each make one decision, and they need to decide simultaneously. These so-called normal form games can be represented by a table (*payoff matrix*).

Example 3.2 (**Split or Steal**)
Let's consider the game 'Split or Steal', taken from a game show (e.g., `https://www.youtube.com/watch?v=TKaYRH6E36U`, or even better, `https://www.youtube.com/watch?v=S0qjK3TWZE8&t=159s`). There are two players, who can choose among two actions. Depending on the players' actions, they either share the prize P, one of the players gets all, or they both get nothing. The situation can be represented by the table

	Split	Steal
Split	$P/2, P/2$	$0, P$
Steal	$P, 0$	$0, 0$

Here, one of the players (the blue player) chooses a row of the matrix and the other (the red player) chooses a column. The entries of the table represent the payoff of the respective player. Players wish to maximize their realized payoff. In the following, let's assume we work with the slightly adapted matrix

	Split	Steal
Split	$5,000, 5,000$	$-100, 10,000$
Steal	$10,000, 0$	$0, 0$

(This corresponds to a case of $P = 10,000$ and the case that players are really unhappy if they split and the other steals). In this game, it is noteworthy that while splitting might seem to be the social choice, we probably would predict that rational players would steal. The reason is: even if I'm unsure about what the other player is doing, by stealing I am always better off.

Definition 3.3 (Dominated actions)
An action a is weakly dominated if there is some other action a' which always yields at least the payoff of a, irrespective of the co-player's action. Action a is strictly dominated if a' always yields a strictly larger payoff.

Remark 3.4 (On dominated actions)
The notion of dominated actions captures some kind of minimum requirement that we would expect rational players to do: They should never use actions that are never optimal, for no possible belief on what the co-player might be doing. Hence, we should expect rational players to *eliminate* dominated actions. If, after elimination of all dominated actions, there is only one action per player left, we may call the remaining strategy pair a solution of the game. In this sense, the above 'Split and Steal' can be solved,

and the solution is for both players to steal (even though that outcome leaves everyone worse off on a collective level). However, not all games can be solved in that sense.

Remark 3.5 (A Split and Steal game with remorse)
Consider a version of the above Split and Steal where the row-player would have a bad conscience when stealing against a splitting opponent. In that outcome, the row-player would experience some emotional cost of 8,000 EUR. Then the resulting payoff matrix is

	Split	Steal
Split	$5,000, 5,000$	$-100, 10,000$
Steal	$2,000, -100$	$0, 0$

Now, the row-player is actually open to splitting the amount (provided the co-player goes along). Nevertheless, if both players can be assumed to be rational, there are still reasons to believe they would both steal eventually. The reason is: For the column-player, splitting is still a dominated action; hence that player will certainly prefer to steal. However, if the row-player is aware of its opponent's rationality, the row-player will anticipate the opponent's stealing. Given that belief, it also becomes better for the row-player to steal.

Remark 3.6 (Iterated elimination of dominated actions)
The above procedure to arrive at a prediction is called iterated elimination of dominated actions. Here, we first eliminate actions that are dominated for either of the two players. In a next step, we consider the reduced game that only contains actions that haven't been eliminated yet. Then we also eliminate all actions that are dominated in this reduced game. We iterate this process until no action can be eliminated anymore. If in the end, both players only have a single action, then the game is solvable by iterated elimination of dominated actions (such games are also called 'dominance-solvable'). The above Split and Steal with Remorse is dominance-solvable. However, we note that the required reasoning on part of the players is now more complex. To arrive at this solution, not only does the row-player need to be rational (try to maximize her own payoff). She also needs to believe that her opponent is rational, and she needs to use that belief to correctly anticipate the opponent's behavior.

Example 3.7 (Traveler's dilemma)
Consider the following scenario: an airline lost two identical pieces of luggage by two different travelers. It makes the following offer: It will ask both travelers independently the value of the luggage (within a range between 180 EUR and 300 EUR). Then it will pay the lower of the two claims. But in case the two claims are different, it will also pay a small reward of $R = 5$ EUR to the traveler who made the smaller claim. Can we solve this game by elimination of dominated actions? Can we solve it by iterated elimination of dominated actions?

Remark 3.8 (Optimal behavior in the stag-hunt game)
Unfortunately, it is fairly easy to come up with games that are not dominance-solvable. One example is

the stag-hunt game with payoff matrix

	Stag	Hare
Stag	10, 10	0, 4
Hare	4, 0	4, 4

Opting for a stag-hunt is not dominated – after all, it is the best thing I can do if I expect my co-player to hunt stag as well. Similarly, hunting hare is not dominated either. Still, some outcomes seem less likely than others. For example, it would be odd to consider (Stag, Hare) a 'solution' (i.e., for the row-player to go for a stag-hunt, while the column-player hunts a hare). Here, at least one of the players would immediately want to deviate from that outcome. In that regard, an outcome like (Hare, Hare) or (Stag, Stag) seems much more reasonable.

Definition 3.9 (Nash equilibrium)
A pair of actions (a_R, a_C) for the two players (one action for the row-player and one for the column-player) is a Nash equilibrium if neither player can improve its payoff by unilaterally deviating.

Theorem 3.10 (Nash (1950))
Any normal-form game has an equilibrium point. (*possibly in mixed strategies*; i.e., for some games, players need to be able to randomize between different actions).

Remark 3.11 (On the status of the Nash equilibrium concept)

- It is fair to say that the Nash equilibrium is the standard concept to 'solve' normal-form games.
- In some situations, human behavior is also fairly well predicted by the Nash equilibrium. For example, empirical papers suggest quite a good agreement for penalty kicks in soccer (Chiappori et al., 2002) or for tennis serves (Walker and Wooders, 2001). [Bonus question: Why do researchers use sports to compare the Nash equilibrium concept to data?]
- However, it is also fairly easy to construct artificial games in which human subjects considerably deviate from the Nash equilibrium predictions. For more on that, see Goeree and Holt (2001).

Remark 3.12 (A summary)
In this class, you should have learned

- what game theory is about (about strategic decision-making when individuals need to take others' possible decisions into account)
- how to formalize a game in a mathematical model (by identifying the players, their actions, the order of moves, the information they have, and the payoffs)
- how to 'solve' that model (we considered two solution concepts: iterated elimination of dominated strategies, and the Nash equilibrium).

4 Further material

Remark 4.1 (**On the intersection of game theory and machine learning / NLP)**
Here are some research directions that I see at the intersection of game theory and machine learning.

- **Using machine learning to do better game theory.** Some people use empirical data and machine learning to explore (*a*) whether the Nash equilibrium is in fact a reasonable solution concept for humans, and (*b*) whether we can come up with a better solution concept. An example of this kind of research is by Zhu et al. (2025). Similarly, there is some research that analyzes the arguments people use when trying to convince each other to cooperate in 'Split or Steal', see for example Turmunkh et al. (2019); even though that one is not particularly ML-focused.

- **Comparing LLM behavior to human behavior.** There is quite some interest in exploring (*a*) how LLMs would act in these kinds of games, and (*b*) to which extent LLM behavior is similar to human behavior. Here, an example is Akata et al. (2025).

- **Using game theory to improve ML algorithms.** To improve their ML algorithms, researchers sometimes design a game where one model's task is to find a solution to a given problem, and the other model's task is to challenge that solution. Here, an example is Kirchner et al. (2024)

- **Multiagent reinforcement learning.** There is a huge literature on multiagent reinforcement learning that is relevant to both game theory and machine learning. See for example, Leibo et al. (2017).

Remark 4.2 (**Further reading)**
If you would like to learn more about game theory, there are some great books out there. For example, if you would like to have a similarly low-level introduction to game theory, you could read Binmore (2007). For mathematical details, there is the classical book by Fudenberg and Tirole (1998) among many others. Finally, for people interested in evolutionary game theory (the kind of game theory we do in our group), a good source is the book of Nowak (2006), or if you want to see even more sophisticated mathematics, the book by Hofbauer and Sigmund (1998).

References

Akata, E., Schulz, L., Coda-Forno, J., Joon Oh, S., Bethge, M., and Schulz, E. (2025). Playing repeated games with large language models. *Nature Human Behaviour*, 9:1380–1390.

Binmore, K. (2007). *Game Theory: A very short introduction*. Oxford University Press.

Chiappori, P.-A., Levitt, S., and Groseclose, T. (2002). Testing mixed-strategy equilibria when players are heterogeneous: The case of penalty kicks in soccer. *American Economic Review*, 92(4):1138–1151.

Fudenberg, D. and Tirole, J. (1998). *Game Theory*. MIT Press, Cambridge, 6th edition.

Goeree, J. K. and Holt, C. A. (2001). Ten little treasures of game theory and ten intuitive contradictions. *American Economic Review*, 91(5):1402–1422.

Hofbauer, J. and Sigmund, K. (1998). *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge, UK.

Kirchner, J. H., Chen, Y., Edwards, H., Leike, J., McAleese, N., and Burda, Y. (2024). Prover-verifier games improve legibility of llm outputs. *arXiv*, page 2407.13692.

Leibo, J. Z., Zambaldi, V., Lanctot, M., Marecki, J., and Graepel, T. (2017). Multi-agent reinforcement learning in sequential social dilemmas. *AAMAS'17*, pages 464–473.

Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences USA*, 36:48–49.

Nowak, M. A. (2006). *Evolutionary dynamics*. Harvard University Press, Cambridge MA.

Turmunkh, U., van den Assem, M. J., and van Dolder, D. (2019). Malleable lies: Communication and cooperation in a high stakes TV game show. *Management Science*, 65:4795–4812.

Walker, M. and Wooders, J. (2001). Minimax play at Wimbledon. *American Economic Review*, 91:1521–1538.

Zhu, J., Peterson, J. C., Enke, B., and Griffiths, T. L. (2025). Capturing the complexity of human strategic decision-making with machine learning. *Nature Human Behaviour*, 9:2114–2120.