Agent-based Systems

Paolo Turrini

🏠 www.dcs.warwick.ac.uk/~pturrini ✉ p.turrini@warwick.ac.uk

Strategies

The (tentative!) plan

- **Logical Agents [Week 1-2]**
  - Knowledge, Preferences, Strategies and how to reason.
- **Decision Theory [Week 3]**
  - Probabilistic Beliefs and Expected Utility.
- **Game Theory [Week 4-5]**
  - Extensive Games and Opponent Modelling.
- **Learning Agents [Week 6]**
  - Markov Decision Processes, (Multi-Agent) Learning.
- **Collective Decision-Making [Week 7-8]**
  - Cooperation and Social Choice
- **Social Agents [Week 9]**
  - Coalitions, Matching, Social Networks.

The plan for today

- Defining the game (very abstractly):
  - States and possible states
  - Actions and how they connect states
- Strategies and winning conditions (in chess and beyond)
- Group strategies and (again, very abstractly) coalitional power

We start without numbers, we are going to add them later.

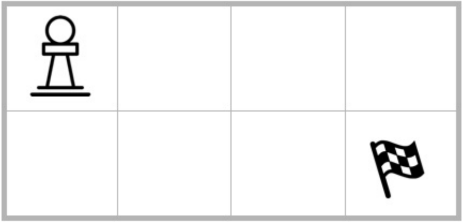
A multi-agent system

- A finite set of **agents**: $N = \{1, 2, \ldots, n\}$
- A set of **worlds** (or **states**): $W = \{w_1, w_2, \ldots\}$
- A distinguished **starting** world (or real world): $\mathbf{w} \in W$
- A finite set of **actions**: $A = \{a_1, a_2, \ldots, a_m\}$
- A set of **terminal worlds** $Z \subseteq W$

The idea:

1. We start from the starting state
2. The agents choose actions
3. The world changes accordingly
4. We stop when we reach a terminal state

A multi-agent system



- ♙ moves
- 🏁 moves too!
- ♙ wants to catch 🏁
- 🏁 wants to catch ♙

Two agents, going Left, Down, Up, Right. Suppose they take turns to move, there is no randomness, hitting the wall results in no movement and the game ends when an agent enters the square occupied by another agent.

Question: How many are the terminal states?

Transitions

How do we formally model the way the world changes?

Let us call $(a_1, a_2, \dots, a_n) \in A^N$ an **action profile**, i.e., the choice of one element of A by each of the agents.

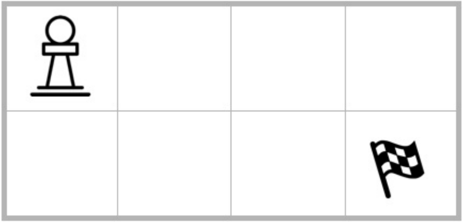
The way the interaction develops is dictated by a **transition function**

$$T : (W \setminus Z) \times A^N \to W$$

which associates a state to each non-terminal state and action profile.

Question: This is a concurrent game, as agents move together.
But turn-based games can be simulated by this general definition. How?

A multi-agent system

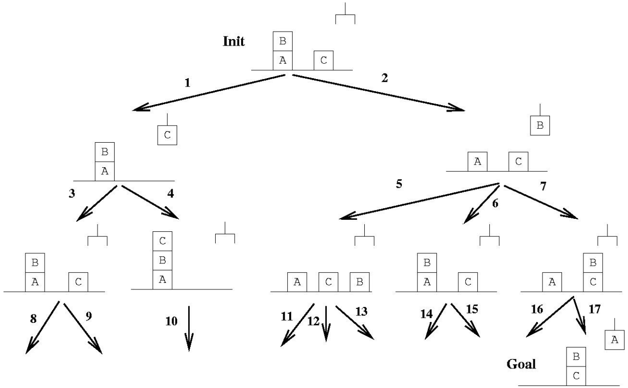


- ♙ moves
- 🏁 moves too!
- ♙ wants to catch 🏁
- 🏁 wants to catch ♙

Two agents, going Left, Down, Up, Right. Suppose they move simultaneously, there is no randomness, hitting the wall results in no movement and the game ends when an agent enters the square occupied by another agent.

Question: Any noticeable differences wrt the previous scenario?

Objectives



Agents are **goal-oriented**: they want to achieve some terminal states.

Objectives

Objectives or goals are a fundamental ingredient of a decision-making agent.
Interesting in single agent contexts, but even more in multi-agent ones:

- competitive: "I want what you don't want"
- cooperative: "I want what you want"
- partially cooperative: "I don't mind what you want"
- partially cooperative (v2): "I'm happy with what you want, but even happier with what you don't"
- instrumental: "I want x, because with x I can get y"
- multi-agent: "I want that you want x"

Goals are often not reducible to win-lose situations.
But let us start with the 'simple' cases.

Winning conditions

Let Z be the set of terminal states.

A **winning set** for agent i is a subset of Z.

The idea: they are the goal of agent i.

An interaction is **strictly competitive** if the set of all winning sets is a partition of Z.

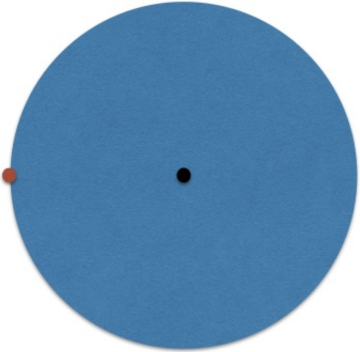
The idea: agents either win or lose.

It is **weakly competitive** if they only partition a strict subset of Z.

The idea: agents can draw.

Single actions are typically not enough to win a game.
We need something more: strategies.

What is a strategy?



The blue circle is a lake. The black point is you, on a boat. The red point is a brute, who wants to catch you. The brute can't swim but moves quicker than you: if you travel a radius, the brute travels half a circumference. On land, you are quicker.

How do you escape? [1]

[1] Credit for this example goes to Chess GM Mihai Suba.

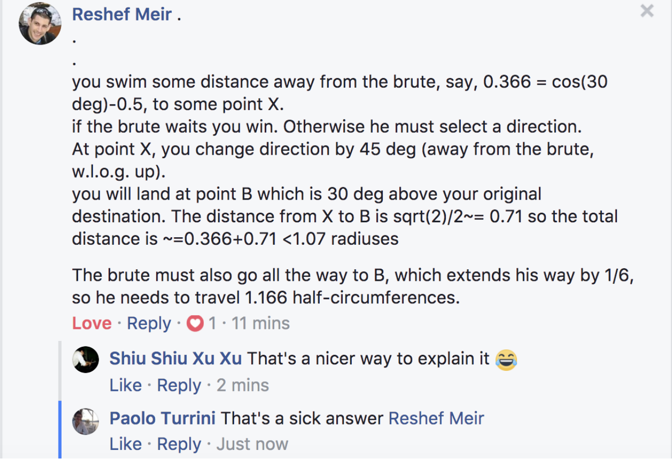
What is a strategy?

"I will not spoil your own pleasure in discovering the solution. I want to emphasise that here you are not supposed to find a move but a strategy, that is a succession of moves, each one depending on the sum total of your own and your opponent's previous moves, and leading to a clear result."

Mihai Suba,
Dynamic Chess Strategy, 1991

Mihai Suba

My Facebook friends



Reshef Meir .

.

.

you swim some distance away from the brute, say, 0.366 = cos(30 deg)-0.5, to some point X.
if the brute waits you win. Otherwise he must select a direction.
At point X, you change direction by 45 deg (away from the brute, w.l.o.g. up).
you will land at point B which is 30 deg above your original destination. The distance from X to B is sqrt(2)/2~= 0.71 so the total distance is ~=0.366+0.71 <1.07 radiuses

The brute must also go all the way to B, which extends his way by 1/6, so he needs to travel 1.166 half-circumferences.

Love · Reply · ❤ 1 · 11 mins

Shiu Shiu Xu Xu That's a nicer way to explain it 😂

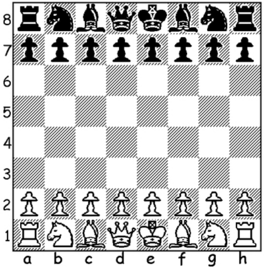
Like · Reply · 2 mins

Paolo Turrini That's a sick answer Reshef Meir

Like · Reply · Just now

Chess

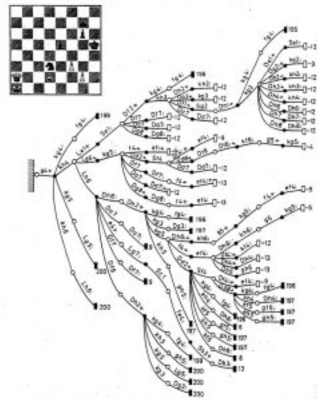
- Two-player
- Turn-based
- Pieces move according to specific rules
- Game ends when a player captures the opponent's king [2]



[2] This is pretty much the same as using checkmate as terminating condition. My educated guess (I haven't checked) is that this is a more ancient variant of the game, from which the modern version was born.

Chess

- The rules of the game (in particular captures and three-fold repetition) force the game to be finite
- Mathematically, it is a representable as a finite tree, where players (White and Black) take turns to move



Chess

- Our set of states is the set of all possible board positions. The starting state is the standard starting board position.
- Notice that a board position might be reached in different ways. We call these ways **histories**.

Question: is the set of histories finite?

Chess histories, formally

Definition (Chess Histories)

A **history** is a sequence $(x_0, x_1, \ldots, x_K)$ such that

- x_0 is the opening board position
- For each even integer k with $0 \leqslant k < K$, going from position x_k to position x_{k+1} can be accomplished by a single legal move by White.
- For each odd integer k with $0 \leqslant k < K$, going from position x_k to position x_{k+1} can be accomplished by a single legal move by Black.

Chess strategies

Suppose we want to construct a computer program to play chess.

This program will have to take a decision at each history, where its assigned role (White or Black) can move.

We call this decision a **strategy**.

Chess strategies, formally

Definition (Strategies)

A **strategy for White** is a function σ_W that associates to each history $(x_0, x_1, \ldots, x_K)$, with K even, a board position x_{K+1} reachable by White with a legal move.

A **strategy for Black** is a function σ_B that associates to each history $(x_0, x_1, \ldots, x_K)$, with K odd, a board position x_{K+1} reachable by Black with a legal move.

A **play of the game** is a pair of strategies (σ_W, σ_B).

Winning strategies

A play of the game ends either in:

- a victory for White
- a victory for Black
- a draw

Definition (Winning strategies)

A strategy for White is called **winning** if, no matter the strategy chosen by Black, it guarantees a win for White.
Formally:

σ_W is **winnng for White** if and only if $\forall \sigma'_B, (\sigma_W, \sigma'_B)$ wins for White.

For Black, the definition of winning strategy is the same
(but with reversed names!)

At-least-drawing stratgies

Definition (At-least-drawing strategies)

A strategy for White is called **at-least-drawing** if, no matter the strategy chosen by Black, it guarantees a win or a draw for White.
Formally:

σ_W is **at-least-drawing for White** if and only if
$\forall \sigma'_B, (\sigma_W, \sigma'_B)$ wins for White or draws.

For Black, the definition of at-least-drawing strategy is the same
(but with reversed names!)

An ancient result

Theorem (Zermelo (1913), von Neumann (1928))

In Chess, one of the following must be true:

- *White has a winning strategy*
- *Black has a winning strategy*
- *Both players have an at-least-drawing strategy*

Over 100 years and we still don't know which one of them is true!

Proof (1/3)

Proof.

Let us first recall that the game is finite, i.e., there is a natural number K such that every play of the game concludes after at most $2K$ rounds, (K turns by White, K by Black).

Proof (1/3)

Proof.

Let us first recall that the game is finite, i.e., there is a natural number K such that every play of the game concludes after at most 2K rounds, (K turns by White, K by Black).

Assume there are exactly 2K turns in every play of the game. Notice that if some plays are shorter, we can simply continue them by adding a do nothing move, preserving the result (so, for instance if we extend a play where White wins, we keep track of this).

Proof (2/3)

Proof.

For every k with $1 \leqslant k \leqslant K$ denote:

Proof (2/3)

Proof.

For every k with $1 \leqslant k \leqslant K$ denote:

- a_k the move implemented by White at their turn.
- b_k the move implemented by Black at their turn.

Proof (2/3)

Proof.

For every k with $1 \leqslant k \leqslant K$ denote:

- a_k the move implemented by White at their turn.
- b_k the move implemented by Black at their turn.

Denote W the fact that White wins (after $2K$ turns), $\neg W$ the fact that White does not.

Proof (2/3)

Proof.

For every k with $1 \leqslant k \leqslant K$ denote:

- a_k the move implemented by White at their turn.
- b_k the move implemented by Black at their turn.

Denote W the fact that White wins (after $2K$ turns), $\neg W$ the fact that White does not.

But then, the fact that White has a winning strategy can be written as:

$$\exists a_1 \forall b_1 \exists a_2 \forall b_2 \ldots \exists a_K \forall b_K (W)$$

Proof (3/3)

Proof.

So, the fact that White has not a winning strategy can be written as:

$$\neg\exists a_1 \forall b_1 \exists a_2 \forall b_2 \ldots \exists a_K \forall b_K (W)$$

Proof (3/3)

Proof.

So, the fact that White has not a winning strategy can be written as:

$$\neg\exists a_1 \forall b_1 \exists a_2 \forall b_2 \ldots \exists a_K \forall b_K (W)$$

This, using first-order logic, is equivalent to:

$$\forall a_1 \exists b_1 \forall a_2 \exists b_2 \ldots \forall a_K \exists b_K (\neg W)$$

Proof (3/3)

Proof.

So, the fact that White has not a winning strategy can be written as:

$$\neg\exists a_1 \forall b_1 \exists a_2 \forall b_2 \ldots \exists a_K \forall b_K (W)$$

This, using first-order logic, is equivalent to:

$$\forall a_1 \exists b_1 \forall a_2 \exists b_2 \ldots \forall a_K \exists b_K (\neg W)$$

But this says that Black is guaranteed at least a draw!

Proof (3/3)

Proof.

So, the fact that White has not a winning strategy can be written as:

$$\neg \exists a_1 \forall b_1 \exists a_2 \forall b_2 \ldots \exists a_K \forall b_K (W)$$

This, using first-order logic, is equivalent to:

$$\forall a_1 \exists b_1 \forall a_2 \exists b_2 \ldots \forall a_K \exists b_K (\neg W)$$

But this says that Black is guaranteed at least a draw!

We can do exactly the same for Black.

Proof (3/3)

Proof.

So, the fact that White has not a winning strategy can be written as:

$$\neg\exists a_1 \forall b_1 \exists a_2 \forall b_2 \ldots \exists a_K \forall b_K (W)$$

This, using first-order logic, is equivalent to:

$$\forall a_1 \exists b_1 \forall a_2 \exists b_2 \ldots \forall a_K \exists b_K (\neg W)$$

But this says that Black is guaranteed at least a draw!

We can do exactly the same for Black.
Therefore, one of the three alternatives must hold.

□

Solving a game vs beating the World Champion

Solving a game = finding the **objectively** best continuation from the start.

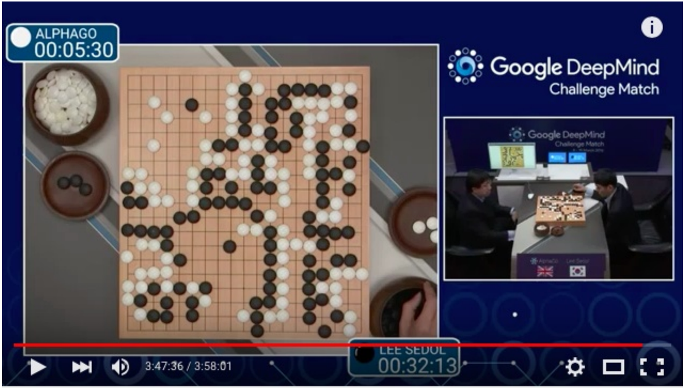
Solving a position = solving the game starting from there.

AI is very good at some games. But is it perfect?

Solving a game vs beating the World Champion

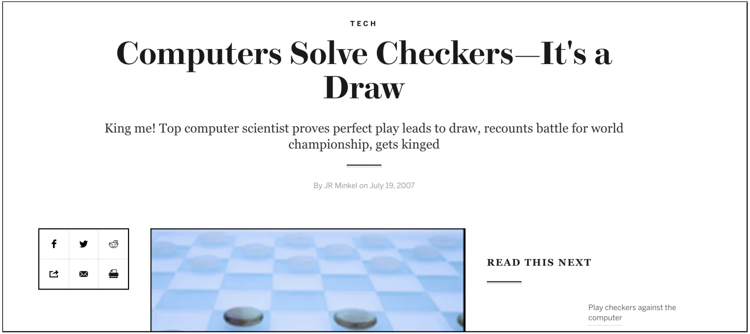
Chess is **not** a solved game. Even if DeepBlue has beaten Kasparov.

Solving a game vs beating the World Champion



Go is **not** a solved game. Even if AlphaGo has beaten Lee Sedol.

Solving a game vs beating the World Champion



TECH

Computers Solve Checkers—It's a Draw

King me! Top computer scientist proves perfect play leads to draw, recounts battle for world championship, gets kinged

By JR Minkel on July 19, 2007

READ THIS NEXT

Play checkers against the computer

Checkers **is** a solved game. It's a draw.

Solving a game vs beating the World Champion

Chess is not solved, but some chess endgames are.

Chance it gets solved any soon?

https://arxiv.org/abs/1712.01815

Group Strategies

Long-range ability

Recall that a transition function associates to each (non-terminal) state and action profile a successor state.

$$T : (W \setminus Z) \times A^N \to W$$

Now we take a more general stance and analyse what agents can achieve as an outcome, together.

Power and the EU

Example

The treaty of Rome (1958-1973) established the European Economic Community. According to Article 148 of the Treaty, acts of the Council (one of the main legislative institutions) required for their adoption:

- 12 votes (if the act was proposed by the Commission), or
- 12 votes by at least 4 member states (if not).

The values above refer to the EU-6, the founding member states. The treaty allocated the votes as follows:

- 4 votes: France, Germany, Italy;
- 2 votes: Belgium, The Netherlands;
- 1 vote: Luxembourg.

What we care about is what “coalitions” can achieve.

Effectivity Functions

An **effectivity function** is a function

$$E : 2^N \to 2^{2^W}$$

which associates to each group of agents a set of sets of worlds. Intuitively, $X \in E(C)$ means that group of agents C can achieve worlds X.

Outcome monotonicity

We require E to be **outcome monotonic**:

$$X \in E(C) \text{ and } X \subseteq Y \text{ implies } Y \in E(C)$$

"if a coalition can force a set of outcomes, then they can force each superset"

Does it make sense to you?
What other properties do you think are natural?

Outcome monotonicity

We require E to be **outcome monotonic**:

$$X \in E(C) \text{ and } X \subseteq Y \text{ implies } Y \in E(C)$$

"if a coalition can force a set of outcomes, then they can force each superset"

Does it make sense to you?
What other properties do you think are natural?

Exercise: How would you model the EU-6 scenario with effectivity functions?

Power and the EU

If we model a possible world as a possible outcome of a vote...

$E(France, Germany, Italy) = \{\{w\} \mid w \in W\}\}$ = everything

$E(Belgium, TheNetherlands, Luxembourg, France) = \{W\}$ = nothing

$E(C) = E(C \cup \{\text{Luxembourg}\})$ = Luxembourg never counts for anything

Zermelo's Theorem again

If we model $\{win_i\}$ as the set of winning histories for player i:

Zermelo's Theorem again

If we model $\{win_i\}$ as the set of winning histories for player i:
Exercise: Tell me what is coming next :)

Zermelo's Theorem again

If we model $\{win_i\}$ as the set of winning histories for player i:

Exercise: Tell me what is coming next :)

Theorem

In Chess, one of the following must be true:

- $\{win_W\} \in E(White)$
- $\{win_B\} \in E(Black)$
- $W \setminus \{win_W\} \setminus \{win_B\} \in (E(Black) \cap E(White))$

What we have seen

- States, actions, goals
- Strategies and winning conditions
- Zermelo's Theorem
- Effectivity Functions

What Next?

In Chess, we might have a strategy to win or to draw.

But, if we don't know it, there is not so much we can do....

We are going to be talking about knowledge
(e.g., knowing that you can win vs. knowing how you can win)