LRP Week 13-1 Chains of Reasoning, and "Thinking"

1 | Which Things Count as Reasoning? (AlphaGo edition)

1.1 | *Categorizing Examples*

For each of the examples, discuss with your group and decide

1. Whether you think the example counts as a case of reasoning (yes/no/not sure)
2. What you think makes it the case that or explains why the example counts / doesn't count as a case of reasoning.

Examples

(1) **Maximization Game (Human)** You try to win the maximization game. Is it reasoning when you figure out the best route?

(2) **Maximization Game (Algorithmn)** You write a computer program to brute-force calculate the best route in the maximization game. Is it reasoning when this algorithm runs?

(3) **LLM**: You train an LLM on a bunch of your writing. Then you give it a paragraph you wrote about the ethical implications of processed food, and ask it to generate the most likely next words you would have written for the next paragraph. Is it reasoning when the computer implements the probabilistic calculations to generate the most likely words?

(3a) If you think (3) counts as reasoning, is the LLM reasoning *about processed food?* If not, what is it reasoning about?

(4) **AlphaGo**:

1. Given a bunch of examples of how Go masters played go, a model is trained to generate the "most likely next move" given a state of play. This is like **LLM**
2. Then the model generates a string of these "most likely" moves, and there is a scoring rule that determines which is most likely to end in victory. This is like **Maximization Game (Algorithm)**.

Is AlphaGo reasoning about the best way to play Go?

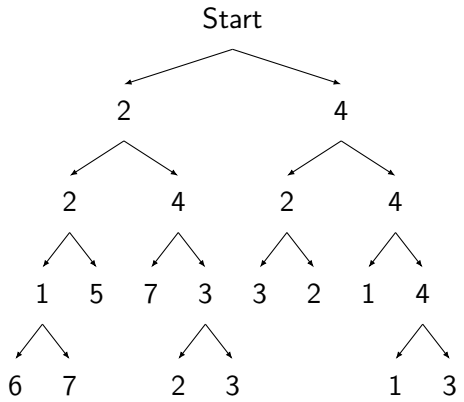
If we get too daunted by the question of whether something (like AI) counts as reasoning, we could instead ask what the stakes are. For AI reasoning, some reasons we might care about what AI can do:

1. **Accuracy**: is AI good enough at getting to the correct answers in problem types we do know how to solve, but don't want to always solve ourselves?
2. **Generalization**: Can AI extend patterns in ways that get the underlying structure of problem types we know how to solve, or is it doing "nothing more than" local pattern-recognition on input data?
3. **Novelty**: can AI help solve problems in *new* ways, ways that humans haven't even thought of?

In general, when faced with a daunting, philosophical, "what is X?" question, it can be helpful to

1. Not try to define X in terms of other words that might be just as hard to define (like "thinking" or "consciousness", for "reasoning")
2. Instead, ask *why* we care about figuring out "what is X," and see if that can help focus our attention.

The Maximization Game



At each juncture, select a direction. Collect as many points before you get to a dead end.