LRP Week 12-1 The "Reasoning" in LRP

1 | What is Reasoning?

Some things we've called "reasoning" this term:

1. Deductive arguments: lists of statements (premises and a conclusion) where the premise is supposed to guarantee the conclusion.
2. Inductive arguments: lists of statements (premises and a conclusion) where the premise is supposed to make the conclusion more likely.
3. Using and improving instances of logic machines.
4. Argument mapping and critique.
5. Figuring out what is most probable by using tools of probability like degrees of confidence, the ratio formula, and theorems that follow from them.
6. Updating your opinions using the norm of conditionalization.
7. Figuring out what is most probable by asking what is evidence for what.

But these are just examples! What *is* reasoning?

1.1 | An Example

Example One

We have three cups:

1. One has two blue m&ms.
2. One has two yellow m&ms.
3. One has a blue m&m and a yellow m&m.

We randomly select a cup. Then we randomly take out an m&m. It is [blue/yellow]. What is the probability that the second m&m we take out is also [blue/yellow]?

As we solve this problem, let's consider: if we think we're *reasoning* when we're solving this problem, what are we doing that makes it reasoning?

Let E = the m&m we took out is [blue/yellow].
Let H = The remaining m&m in the cup is also [blue/yellow].

Recall:

1. By the Evidence Lemma: E is evidence for H if $Pr(E \mid H) > Pr(E \mid \neg H)$.
2. By Bayes' Theorem:

$$\Pr(H \mid E) = \Pr(H)\frac{\Pr(E \mid H)}{\Pr(E)}.$$

2 | Exercise: Which Things Count as Reasoning?

2.1 | *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) We use tools of probability to answer the question in Example 1.

(2) I give Adam a list of probability puzzles and the correct answers to them. Then I give him a problem from the list, and he looks up the correct answer and gives it to me.

(3) I know that if the bus was late, then Ben won't catch his train, and if Ben doesn't catch his train, he won't make it to class on time. Ben isn't in class on time, and I conclude because of that that the bus was late.

(4) I give you a spreadsheet that implements Bayes' Theorem: if you give it $\Pr(E)$, $\Pr(E \mid H)$, and $Pr(H)$, it calculates $Pr(H \mid E)$. (Consider both: Is the spreadsheet reasoning when it does the calculation? Are you reasoning when you plug in the numbers?)

(5) We ask Claude to solve the problem in Example 1. It outputs a list of words and symbols that approximates what we said.

2.2 | *Proposing Features*

Based on the answers you gave above and your reasoning, propose three features of reasoning. Put each feature in one (or both) of the two formats below:

1. When [insert feature] is present, reasoning is occurring.
2. When [insert feature] is *not* present, reasoning is *not* occurring.

Exit Ticket: Write on the card the feature you think is *most* important to reasoning, of the three (+ your name and the date)

3 | Argument Schemas for LLM Reasoning

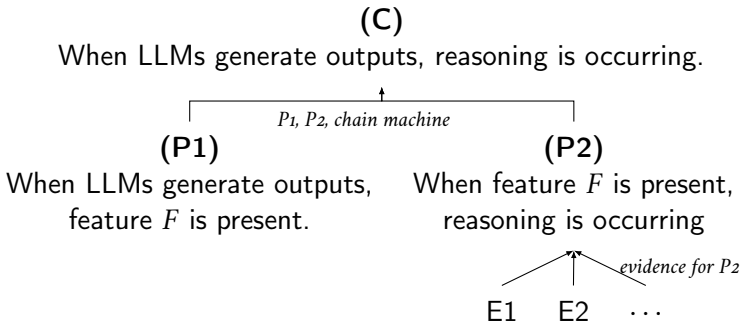
Schema 1: *For* LLM Reasoning

Let F be some feature that we think makes something reasoning.

(P1) When LLMs generate outputs, feature F is present.

(P2) When feature F is present, reasoning is occurring.

(C) When LLMs generate outputs, reasoning is occurring.



Example:

(P1) When LLMs generate outputs, they use a chain of inference.

(P2) When chains of inference are present, reasoning is occurring (e.g. Example 3)

(C) When LLMs generate outputs, reasoning is occurring.

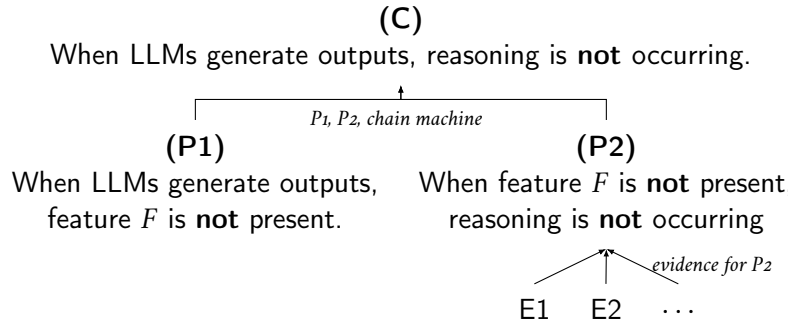
Schema 2: *Against* LLM Reasoning

Let F be some feature that we think makes something reasoning.

(P1) When LLMs generate outputs, feature F is **not** present.

(P2) When feature F is **not** present, reasoning is **not** occurring.

(C) When LLMs generate outputs, reasoning is **not** occurring.



Example:

(P1) When LLMs generate outputs, they do not show robust generalizability.

(P2) When robust generalizability is not present, reasoning is not occurring (e.g. Example 2)

(C) When LLMs generate outputs, reasoning is not occurring.

4 | Chatbot Probability

Let's give Anthropic's Claude Sonnet 4.5 Model the m&ms puzzle. Given our discussion, do we think it is reasoning?

5 | Reasoning and Autonomy

Aylsworth and Castro say things about "autonomy" that make it sound like reasoning is important: They say autonomy involves a capacity to set and pursue your own goals, and an ability to critically assess your values, and to reflect on your conception of the good life and revise it.

Moreover, *reasoning* seems to be part of why A&C think humanities papers are particularly important assignments:

> By writing your own humanities papers and engaging with your professor's feedback, you are honing your ability to weigh different commitments, to assess the strength of competing arguments, and to render your worldview coherent, skills that are definitive of, or, at the very least, partially constitutive of autonomy (19-20).

So it seems like they're saying that part of cultivating your autonomy is developing and preserving the ability to *reason* about which goals matter to you, and how to pursue those goals.

This raises some questions:

1. Suppose you think you should respect your own autonomy. Does this require always doing things in a way where you're reasoning *yourself* about them?
2. How much AI use / assistance is compatible with still counting as "reasoning yourself" about things?
3. Does the answer to the above question depend on whether AI itself counts as doing reasoning?

Bonus: Reasoning about What?

Even if you think AI is reasoning, it might be relevant *what* the AI is reasoning about. In a sense, whether you ask an LLM a question, the LLM is actually giving an answer to a different question:

> *What is the answer to [the question] that is most likely to be correct according to the training data?*

This is like if an interviewer asks "what is your greatest weakness?" and you think of the response that answers the question:

> *What is the answer to "what is your greatest weakness" that is most likely to get me the job?*

So, is it enough that the LLM answers this different question?