



## AI explains itself

### Description

I'm catching up with Yuval Harari's latest book *Nexus: A Brief History of Information Networks from the Stone Age to AI*. He alerts readers to the challenge posed by the spread of opaque algorithms that are increasingly responsible for deciding on our behalf: bank credit, purchase choices, health, governance. In the face of this obfuscation of accountability he advocates a "new human right: the right to an explanation" (331). This is one of the challenges of AI, to "explain itself."

OpenAI has introduced a variant of ChatGPT that generates responses to prompts that deal with logical inference, and that attempts to explain itself in the process. On the [ChatGPT-o1](#) website they claim "We trained these models to spend more time thinking through problems before they respond, much like a person would. Through training, they learn to refine their thinking process, try different strategies, and recognize their mistakes." I put ChatGPT-o1 to the test with a simple sequencing task I introduced in an earlier post: [Robot probes city grid](#).

I told ChatGPT-o1 about my earlier demonstration (using ChatGPT-4) in which I uploaded a couple of images and asked the model to show how best to move objects around to transition from the configuration in the first to the second scene. I asked if ChatGPT-o1 could infer from a copy-paste of that earlier dialogue what those images were, and how to correct the sequence of moves. The platform gave an explanation of the steps it was taking in real time. Here's what is said:

**Taking a closer look:** I'm analyzing a blog post describing an exercise where images were uploaded to ChatGPT. The goal was to identify and correct an ambiguous sequence of moves to transition between images.

**Analyzing the image:** I'm piecing together the user's text to understand the configuration of objects in the initial scene and to conceive a strategic plan for moving objects to the subsequent scene.

**Adjusting my approach:** I'm gathering insights on rearranging items to clarify ambiguous instructions, ensuring smooth progression with clear, detailed steps.

**Reflecting on AI evolution:** Interestingly enough, while aligning with Terry Winograd's SHRLUDU program's conceptual underpinnings in natural language processing, the current chatGPT-based tasks diverge in using neural network architectures instead of rule-based frameworks for AI.

**Documenting progress:** OK, let me see. The user's updating sequence involves moving both containers, temporarily placing them for the swap. The teapot ends up inside the jug, creating a larger container. This progression steps through the rearrangement.

**Revisiting the sequence:** First, I gently lift the miniature teapot from the jug and set it on the table. Next, I pick up the jug and place it on the coaster. Finally, I position the miniature teapot back into the jug.

**Rearranging kitchenware:** OK, let me see. I'm gathering info on placing the teapot inside the jug. Using one hand, gently lift the teapot from the jug and set it down. Then, place the jug where the teapot was, and finally nestle the teapot inside the jug.

**Revisiting instructions:** OK, let me see. The initial image shows the jug and teapot in specific places. The goal is to swap their positions with only one hand. The final image likely shows the teapot inside the jug, after adjusting the sequence for clarity.

**Shifting priorities:** I'm tackling the sequence of moves to move a jug and miniature teapot, considering the constraints and initial positions shown in the images.

**Rethinking the sequence:** I'm beginning to rethink the order of picking up the jug and teapot. Adapting to using one hand, I'll incorporate suggestions to improve the sequence, ensuring the teapot is effectively positioned inside the jug.

**Placing the miniature teapot:** OK, let me see. The first step involves lifting the miniature teapot with one hand and putting it into the jug. If the jug needs shifting to the right spot, I'll move it next.

**Refining the approach:** I'm thinking through a more efficient method to swap the jug and miniature teapot, ensuring the teapot stays safe on top during each movement to prevent any damage.

**Rearranging items:** I'm simplifying the process by moving the jug first, then the teapot, minimizing movement and using just one hand.

The explanation concluded that the process took 26 seconds. (The documentation for the platform explains that these explanations are summaries of more detailed but less legible accounts of its processes, presumably available for further processing.)

The platform then presented two methods for moving the objects to their various positions. These were correct, though I hadn't introduced the constraint that the top surfaces of the boxes were not large enough to support both the teapot and the jug next to each other. ChatGPT-o1 does not as yet support images. So, it wasn't able to make that inference.

I then explained that the original demonstration had photographs, showing a table with two small boxes. The top surface of the boxes is such that they can each easily support a jug and a miniature teapot, though not both together. The boxes are of different heights, but that just serves to identify them.

I explained that in the first photograph the teapot is on the taller box and the jug is on the shorter box. The second photograph shows the jug now positioned on the taller box with the teapot positioned elegantly atop the jug and nestled into its opening.

I asked what sequence of steps would get me from the arrangement in the first to the second photograph. After a similar account of its own step-by-step inference, it came up with a detailed plan of action.

### **Initial Configuration:**

- **Taller Box:** The miniature teapot is placed here.
- **Shorter Box:** The jug is placed here.
- **Constraints:**
  - Each box can support either the jug or the teapot, but not both simultaneously.
  - You have only one hand free to move objects.

### **Goal Configuration:**

- **Taller Box:** The jug is now placed here.
  - **Jug:** The teapot is elegantly positioned atop the jug, nestled into its opening.
- 

### **Step-by-Step Sequence:**

#### **Place the Jug (with Teapot Atop) onto the Taller Box:**

#### **Pick Up the Teapot from the Taller Box:**

Use your free hand to lift the miniature teapot off the taller box.

#### **Place the Teapot Atop the Jug on the Shorter Box:**

Carefully place the teapot onto the jug that is on the shorter box, nestling it into the jug's opening.

*Note:* Ensure the teapot is securely positioned so it won't fall when you move the jug.

#### **Lift the Jug (with Teapot Atop) from the Shorter Box:**

With your one hand, grasp the jug firmly while keeping the teapot balanced atop it.

Lift both together as a single unit.

---

### Place the Jug (with Teapot Atop) onto the Taller Box:

Move the jug with the teapot atop from the shorter box to the taller box.

Gently set it down to complete the desired arrangement.

---

### Final Result:

- The jug, with the teapot elegantly nestled into its opening, is now positioned on the taller box.
- You've successfully transitioned from the initial to the final arrangement using only one hand.

### Additional Notes:

- **Safety First:** When lifting the jug with the teapot atop, move slowly to maintain balance and prevent the teapot from toppling.
  - **One-Handed Constraint:** All steps are designed to be performed with just one hand, adhering to your constraint.
  - **Object Support Limitations:** At no point are both the jug and teapot placed side by side on the same box, respecting the boxes' support limitations.
- 

### Conclusion

By following these steps, you efficiently rearrange the items to match the second photograph's arrangement, all while using only one hand and respecting the physical constraints of the boxes and objects.

---

### Reflection on the Process:

This exercise mirrors the kind of spatial reasoning and planning found in early AI programs like Terry Winograd's SHRDLU. While modern AI models like ChatGPT generate responses based on learned patterns, incorporating logical planning steps enhances the clarity and effectiveness of instructions for physical tasks.

---

Enjoy your elegantly arranged tabletop display! If you have any more tasks or need further clarification, feel free to ask.

The task was a simply one, and both the explanation and the process are disproportionately lengthy, though I could easily have asked for brief summaries. It's interesting how the platform introduced value-laden warnings about the precariousness of moving objects about whose geometry does not normally suggest they will support one another, e.g. teapots on top of jugs. I assume we can attribute that sensitivity to the patterns of semantic associations evident in the training data, i.e. the corpus of all

---

human-authored text that supported the training of the large language model.

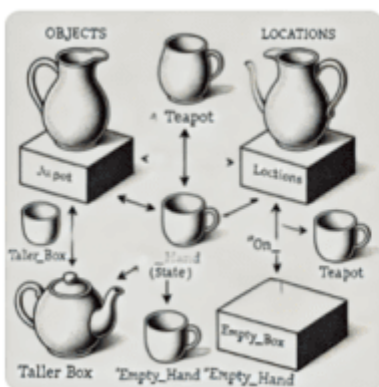
I subsequently asked the system to write down its description of the blocks world and the logic of the various interactions as logical predicates, in the Prolog language. It did this successfully, suggesting the possibility of formalising objects and relationships as derived from images and discursive text, and using that as a way of extending the scope of queries beyond the immediate context. E.g. [ChatGPT finished the paragraph for me] the system could infer rules about object placement or relationships in new scenarios, such as reasoning about how objects might be arranged on a shelf or table based on prior knowledge. This ability opens up avenues for broader context-aware reasoning, enabling the system to generalize from specific tasks (e.g., how to move a teapot from one box to another) to more abstract, higher-level reasoning (e.g., how to organize objects in a room or how to rearrange items to optimize space). In urban contexts, such formalization could be applied to infer complex spatial relationships and interactions, such as optimizing traffic flows, organizing city resources, or planning urban layouts, all derived from similar logical principles and real-world data.

## Reference

- Harari, Y. N. (2024), *Nexus: A Brief History of Information Networks from the Stone Age to Ai*, New York: Penguin.

## Note

Out of curiosity I switched to a version of ChatGPT that supports image generation (ChatGPT-4) and asked it to generate some illustrative images from the output of ChatGPT-o1. Here are some attempts.



## Category

1. Artificial Intelligence

## Tags

1. AI planning
2. blocks world
3. inference
4. logic

## Date Created

September 21, 2024

## Author

rcoyne99

*default watermark*