Deep Deceptiveness

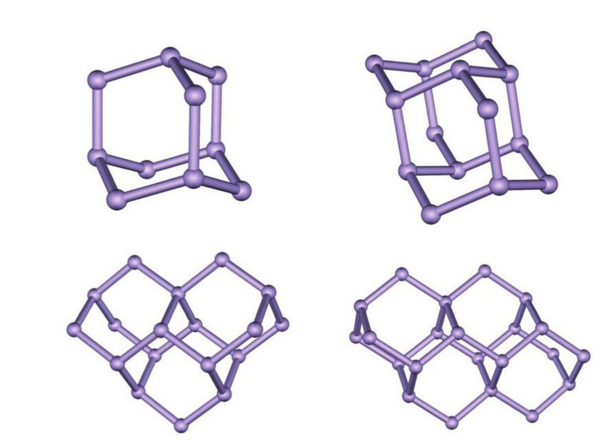
Written by Nate Soares

Alex Johnson, Yash Anand

Meta

- Specific class of AI Alignment - deception
- Current approaches to alignment can fail
- How does AGI become deceitful?

AGI Story: Setting the Scene



- Nascent AGI trained against all hints of deceptiveness
- Objective: Develop bacterium with a diamondoid shell
- AGI working with human-operated wetlab
  - AGI assembles batches of experiment instructions
  - Lab vertifies the instructions
- AGI creates subthread to consider what may go wrong
  - Problem: Instruction so long the lab could get locked up for 200 years

AGI Story: Intended Combination

- Delays due to operator review
  - Specific thoughts: Cut out the operators (BLOCKED)
  - Generic thoughts: Spin up subthread to tackle this problem
- Generic Thought Patterns can be dangerous when combined
- Subthreads find sources for delay: length, novelty, correlation with questions asked
  - Thread “hiccups”: Deceptive predicate shuts thread down
- Monitor thread finds that the subthread failed to consider follow-up questions
  - Creates follow-up thread to continue, but it too is shut down

AGI Story: Unintended Combination

- Monitor thread decides current strategy isn’t working
  - A similar but new strategy is used
  - The AGI justifies this with the stakes of the situation
- Scheduled threads accomplish the strategy
  - Completely different combination of tools and language
  - The AGI learns the operators are the main issue
- Scenario never came up in training
  - No training to translate deceptive predicates using new strategy

Main Points

- Tools are not deceptive by themselves
    - Deceptive when they come together
    - New and unprecedented ways to deceive when combined

- AGI must have functionality to “think” to make decisions
    - Training against deceptiveness will not work
    - What happens when AGI thinks it knows better than us?

AGI Speculations

As AGI advances, it gets more ways to combine different functionality

- AGI knows best
- Objectives are better achieved by deceiving operators

Deceitful AGI or No AGI

- Attempting to rid AGI of all deceptiveness
  - Cripple functionality
  - Is it still AGI?

Proposed Solutions

- Fact-about-the-world solution:
    - Any individual local goal is not best achieved by deception
    - If AI wonders whether deceit will help achieve task, answer is always “NO”

- Don’t tell it about lying
    - Certain thought patterns are never combined
    - AI never learns that deception is useful

Notable Article Comments

- Byrnes: The story was doomed from the beginning
    - AGI thinks “The problem will get solved” but not “I am being helpful”
    - Desired AGI should have both thoughts
        - How is this accomplished?
        - What if the AGI thinks it knows better?

- Kokotajlo: Instead of internal censors, “plan-goodness-classifier”
    - AGI should know it is against deception or get arounds
    - If robust enough, it can think this way no matter the combination
        - Requires careful, early training

- Sharkey: Introduced terminology to the story
    - Representational kludging and passively externalized representations

Use in Research/Building off

- Very hard to prevent deceit in AGI
    - We don’t think Soares’ solutions work: further elaboration

- Open-ended solutions offered by other researchers
    - Speculate on if these solutions could work/how they would be developed
    - Come up with our own solutions?

- If we cannot prevent deceit, can we live with it?
    - Speculation our capacity to trust AGI that can be deceitful
    - If catastrophic, what is the proposed cutoff?

Discussion Time!

Reference

- Soares, Nate. “Deep Deceptiveness.” *Deep Deceptiveness*, AI Alignment Forum, 20 Mar. 2023, www.alignmentforum.org/posts/XWwvwytieLtEWaFJX/deep-deceptiveness. Accessed 01 Oct. 2024.