Task decomposition for scalable oversight

https://course.aisafetyfundamentals.com/alignment?week=4

Discussion points

0) Terminology
a) Scalable oversight and sandwiching
b) Task decomposition
c) Iterated amplification
d) Chain of thought
e) Least-to-most prompting (two sequential stages)

1) Scalable oversight
a) The role of scalable oversight in the context of AI systems. Examples of tasks that are too complicated for a single human to evaluate

b) What are the potential ways of scalable oversight and alignment techniques evolution in the coming years?

c) What are the downsides of scalable oversight?

2) Task decomposition
a) What is the impact of dividing intricate tasks into smaller, more easily handled parts on AI system performance? Are there any tasks where this approach will fail?

b) How can we identify the necessity for task decomposition (or any particular task decomposition/prompting strategies) to each specific sample of task?

c) When dealing with a particular task, the task decomposition approach would be applied to each individual example. It's possible that the model can handle simple cases on its own. Could applying task decomposition for those simple instances be potentially misleading?

d) How to evaluate the quality of task decomposition? How do algorithms such as iterated amplification can prevent the creation of malicious subtasks?

e) Could certain very complex problems require non-human-like problem solving steps and hence learning by imitating humans might be too limiting?

3) Chain of thought

a) The chain of thought prompting and least-to most prompting seem to rely on the fact that the prompter can verify the correctness of the sub-answers. Can we verify that the sub-answers are even correct if the topic is too advanced for human brains to understand?

b) Regarding teaching models reasoning, surely it had encountered many examples of it during training, so why doesn't it learn it then? It seems to have all the data necessary, but doesn't know how to apply it?

4) Varia

a) What are the possibilities that AI creates a situation where even subtask oversight becomes too complex for humans to understand?

b) Even if AI can explain how it came to a certain conclusion, does this really help with alignment that much? The explanation only needs to be as plausible as to be acceptable to a human, but humans are flawed and can be manipulated or tricked even given the explanation. The explanation just makes it look more credible, which is even more dangerous if the outcome is ultimately incorrect.

c) Would you agree to deploy AI that is not aligned in case alignment tax is too high?

d) One reason why humans can be trusted is that there are repercussions to illegal or immoral behavior. AI systems lack this type of 'skin in the game' and agency - would there be value in having them?