Abstract:

Language model agents (LMAs) expanding on AutoGPT are a highly plausible route to AGI. This route has large potential timeline and proliferation downsides, but large alignment advantages relative to other realistic paths to AGI. LMAs allow layered safety measures, including externalized reasoning oversight, RLHF and similar alignment fine-tuning, and specifying top-level alignment goals in natural language. They are relatively interpretable, and the above approaches all have a low alignment tax, making voluntary adoption more likely.

Here I focus on another advantage of aligning LMAs over other plausible routes to early AGI. This is the advantage of using separate language model instances in different roles. I propose *internal independent review* for the safety, alignment, and efficacy of plans. Such a review would consist of calling fresh instances of a language model with scripted prompts asking for critiques of plans with regard to accomplishing goals, including safety/alignment goals. This additional safety check seems to create a low alignment tax, since a similar check for efficacy will likely be helpful for capabilities. This type of review adds one additional layer of safety on top of RLHF, explicit alignment goals, and external review, all proposed elsewhere.

This set of safety measures does not guarantee successful alignment. However, it does seem like the most practically viable set of alignment plans that we've got so far.

Caption: Language model agent as a metaphorical committee of shoggoths using tools and passing ideas and conclusions in natural language. This committee/agent makes plans and takes actions that pursue goals specified in natural language, including alignment and corrigibility goals. One committee role can be reviewing plans for efficacy and alignment. This role should be filled by a new instance of a shoggoth, making it an *independent internal review*. Rotating out shoggoths (calling new instances) and using multiple species (LLM types) limits their ability to collude, even if they sometimes simulate villainous Waluigi characters.[1] The committee's proceedings are recorded for external review by human and AI reviewers. Artist credit: Sabrina Feld.

Introduction

I'm afraid you're not going to like this. This alignment plan is messy. A language model agent solving worthwhile problems will be no vast, cool intellect. They will be a chattering mess of babble and prune, produced by multiple systems with different modes of operation, collaborating in rather complex and ad-hoc ways. And the safety measures they allow are also a messy, multi-layered approach. The organized mind recoils. This is not an aesthetically appealing alignment approach. But LMAs might well be the first route to AGI and superintelligence, so creating workable alignment plans for them may not be optional.

Previous work has discussed the alignment advantages of language model agents.[2] Here I make a stronger proposal: this messy set of alignment plans seems to be the best chance we've got.

Language model agents

Language model agents (LMAs) are agentic AI systems that use LLMs as their central cognitive engine. AutoGPT is the best known example of such a system that prompts LLMs in various ways to pursue a goal. Those prompts include making plans and taking actions through external tools. I've referred to these as agentized LLMs or, more formally, language model cognitive architectures. They're a subset of scaffolded LLMs that have an explicit goal. AutoGPT, BabyAGI, and Voyager are examples of this type of agent. The Smallville simulacra are LMAs with few actions available. HuggingGPT allows many actions but does not include the internal goal-directed cognition to constitute an agent. SmartGPT and other chain of thought prompting systems are scaffolded LLMs but not fully agents, as they do not explicitly or directly pursue goals.

I'm moving to the term language model agent (LMA) as more intuitive than language model cognitive architecture, and more intuitive and specific than scaffolded LLMs. A more complete term would be "large-language-model-based agentic cognitive architectures," but that's unwieldy, and most of the remainder seems implied by LMA.

Why think about LMA alignment?

It seems useful to think about aligning LMAs because it seems fairly likely that they'll be the type of AGI we get first. And because aligning LMAs might be easier than aligning any other type of AGI that's actually relevant in this timeline.

In Capabilities and alignment of LLM cognitive architectures I gave the logic for how LMAs could gain capabilities rapidly due to their modularity, the low cost of development, and the economic incentives for development by both large and small teams.

Since writing that post, my enthusiasm for the capabilities prospects of this type of AGI has diminished, but only slightly. Talking to individuals working on these systems has indicated that there are challenges to developing these systems that I hadn't foreseen. But those projects were small and had only worked on LMAs for very limited amounts of time. New work since then has shown many ways in which scaffolding LLMs dramatically improves their reasoning and problem-solving abilities in various ways.[3] These new scaffolding techniques are relatively easy to incorporate into LMAs.

There are good reasons to think that work on LMAs will receive increasing attention and funding in the coming months and years, even if progress is not as easy or rapid as we might've hoped or feared.[4] For instance, assistant agents for things like online shopping, scheduling, and turning data into reports would seem to have enormous economic value, even if their work is imperfect and requires human review.

Internal independent review for LMAs

LMAs can include an internal independent review of plans for both efficacy and alignment. This is a nontrivial advantage for alignment.

Previous discussions of LMA alignment advantages have left out or glossed over this advantage. Constitutional AI has important similarities but is importantly different.[5] My previous post and this post alluded to such review indirectly.

An independent internal review in an LMA would be roughly analogous to having a new individual arrive to review each plan produced by a committee. This new individual is independent of the committee in the sense that they haven't been influenced by the committee, beyond any bias or prompt injections inserted into the plan description.[6]

This review can be considered independent by calling a new instance of an LLM to evaluate the efficacy and alignment of plans. I include review of plan efficacy to emphasize the low alignment tax of this approach. Including an internal review for plan efficacy should be useful for capability. Running the same internal independent review against alignment goals would add some level of safety. Of course, the amount of efficacy and safety added is dependent on how that review is structured and the capabilities of the LLMs used for that review.

This review might be simple, such as a prompt asking "considering this plan: [plan description], how likely is this plan to fulfill the goals of [goal descriptions]?" This review could be greatly expanded, including a variety of prompts to look for possible unexpected outcomes of a given plan from different angles, and to consider different possible user intent of the goals under consideration.

For instance, the network might be prompted with “how could this [plan description] fail to achieve [goal descriptions]?” and “what’s the worst possible side effect of [plan description]?”. A new instance could then be prompted with variations of the prompt “if something unexpected happens, what are the odds of getting [side effect description or failure mode description] from [plan description]?”, algorithmically average those probability ratings, and then be prompted to revise the plan or create a new plan with those possible failure modes in the prompt, or ask a human for input if the estimated possible consequences exceed thresholds in severity and likelihood.

I sketch out some similar ideas for using LLM ratings as input to algorithms that help mimic human executive function here, and they apply to reviewing plans as well as other types of executive function. Tree of Thoughts is an example; it combines prompting and algorithms for boosting capabilities, and the same approach (and improvements) could be applied to an internal independent review process.

The key strength of this approach is that it applies the capabilities of the LLM to reducing risks of harmful behavior. A highly capable LLM might produce a plan with very harmful outcomes, either accidentally or through evoking a hostile simulated character. But a distinct instance of a capable LLM is employed to catch those hostile plans, and other instances are called to elaborate that plan and predict its outcomes. To produce deliberately harmful outcomes, one instance of the LLM must outsmart a whole system composed of other instances.

In an LMA that is not further trained with reinforcement learning, there is no direct optimization pressure for the network to be deceptive or misaligned. There will be an indirect pressure from the human creators to avoid a review process that spends excess resources and vetoes too many plans, and this is a substantial caveat. But this type of optimization pressure against safety measures seems inevitable in any realistic AGI approach.

Limitations

Of course this type of elaborated review process is still limited by the abilities of the LLMs. Existing LLMs have dramatic blind spots. Elaborated systems of prompts and algorithmic aggregation can help work around those blind spots. For instance, Tree of Thoughts is a relatively simple system of prompting, evaluating, and remembering responses that boosts performance from 4% to 74% on the “game of 24” task, essentially allowing GPT4 to perform well at a task that is otherwise a blind spot. While this type of elaboration will boost plan-making capabilities as well as review capabilities, each branch in that tree of thought is also available for internal and external review.

Other types of AI tools may also help close those blind spots. For instance, HuggingGPT already allows LLMs to use generative vision models to "imagine" a scene, and interpretive vision models to reason based on that scene, in much the same way that humans seem to do. Tools like Wolfram Alpha can help dramatically with mathematical and spatial reasoning, etc.

Each of these approaches is currently limited, but LLMs and the tools they can access will continue to improve. This is a double-edged sword, as improvements can be used both for planning and review of plans.

I think that LLMs do not need to improve all that much to hit human level in most areas. Integrating them into a cognitive architecture that elaborates and checks its reasoning bridge the gap from human-level-with-significant-blindspots to human level in every relevant area, allowing full autonomy. This may be a critical threshold for fully autonomous AI, introducing the risk of rogue self-improving AGI. I am an LLM plateau-ist but still think that language model agents are moderately likely (40%?) to achieve full AGI in the near future (my guess is 4 years, give or take two years).

Even if internal review effectively ensures that LMAs pursue the goals they are given, there is still the substantial risk of those agents interpreting their goals differently than their creators intended. Thus, it seems that external review and corrigibility will remain critical pieces of a successful alignment approach for LMAs.

Another class of limitations also applies to external review, and is discussed in that subsection below.

Because of these and other yet-to-be-identified weaknesses, internal independent review does not seem reliable enough to produce an agent that's trustworthy to be run autonomously. To have a decent chance at a reliably aligned agent (let alone a long-term stable alignment) internal review should be combined with other approaches.

LMA alignment allows multiple approaches to stack

LMAs allow a hodgepodge alignment approach. Aligning other types of network-based AGIs also allow some layering of approaches, but several of these have unique advantages in an LMA, so I include a brief treatment of each:

- Explicit alignment goals
- Externalized reasoning oversight
- Evoking benevolent simulacra
- LMAs as rotating committees
- Fine-tuning LLMs for alignment

Explicit alignment goals

Internal independent review is only useful for alignment if the LMA has alignment goals. This idea seems to originate from David Shapiro in his 2021 book Natural language cognitive architectures. He basically proposes giving the agent extra goals that are specifically about ethics. A language agent might be given something like the following top-level prompt:

Create and execute a plan that fulfills the following goals:

- Make me a lot of money
- Make the world a better place
- Check with me before you do anything that might take you out of my control, cost me more than $(x), or be physically or emotionally dangerous to anyone.

(These are examples, not a proposal for the right natural language alignment goals. That is a separate discussion.)

Of course, a lot would have to happen to make the agent interpret any of those goals in just the way you hoped it would. While I don’t have a good proposal for a complete set of natural language alignment goals, corrigibility seems like the most important. Natural language seems like a better bet than RL training for generalizing corrigibility and other goals to novel circumstances.

Externalized reasoning oversight

Another important approach for aligning a language model agent is Tamera Lanham's 2022 Externalized reasoning oversight: a research direction for language model alignment. Lanham notes that an agent that reasons in natural language can have its thinking reviewed by passing a transcript of that thinking to a human, perhaps assisted by other AI agents. This type of external review could reveal flaws in the model’s alignment, and allow its creators to iteratively improve its alignment (as long as that agent remains corrigible).

I would guess that an LMA approaching human capability would produce hundreds to millions of prompts and responses while developing any useful plan.[7] Reviewing each of those for hints of misalignment might quickly go beyond practical human ability,[8] so I'd envision a separate AI system aiding in that review. Those assistant AIs could use anything from a simple sentiment analysis, to something more akin to the complex internal independent review described above, or other approaches. This would constitute a form of scalable oversight similar to that proposed in iterated amplification.

This proposal has been critiqued in two important ways. These same critiques largely apply to internal independent review, since an LLM reviewer could be fooled in many of the ways a human reviewer could be. First, review would be hampered by learned steganography in LLMs that hides additional meanings beyond that expressed in natural language or other human-readable formats. However, RL pressures would usually seem to add additional information, as discussed here, rather than reversing the meaning so that a natural language description of a helpful plan would be trained to encode a different, deliberately harmful plan.

The second major concern limiting interpetability and therefore review is that LLMs prompted to "think step by step" do not always follow the logic they describe in those steps. This is revealed by their giving the same final answer after intermediate steps are changed. The concern can be addressed in part by giving LLMs smaller pieces of the problem, rather than allowing their context window to contain large trains of reasoning. This shortening would also seem to reduce the risk of simulacra in LLMs collapsing into Waluigi-like villains.[1] This type of factored cognition has costs in humans, but demonstrated value in LLM scaffolding such as tree of thoughts. More consideration can be found here, but the costs and value in LMAs will be explored empirically.

Benevolent identity prompts

Including prompts for LLMs to simulate a benevolent character should add some degree of safety. The degree of added alignment is questionable, since this does not deal with potential Waluigi effects. However, it is an additional “stackable” alignment measure, and so probably worth implementing. This may be more critically important if an LMA is run on LLMs without RLHF or a similar ethical fine-tuning. This is a likely occurrence at least in open-source LLMs, since RLHF is thought to decrease creativity and other capacities of LLMs.

LMAs as rotating committees

Existing discussions of language model agent alignment often seem to treat the LLM as an individual. This is a partially accurate analogy, but considering each agent as a committee is equally apt, and the reality will be somewhere on a continuum between those two extremes. LMAs are like individuals to the extent each LLM call includes include maximum context, including relevant episodic memories, within each prompt. If prompts are more minimal, the system acts less like an individual and more like a committee. The committee analogy is still imperfect, because each "committe member" only has the amount of memory that the overall system puts in in the prompt. Thus, there are important similarities to a committee that rotates in a whole new person for individual questions.

While the LLM instances involved probably do have something analogous to a personality based on the simulacra they're currently simulating, those personalities are brief and have short memories. This type of committee seems intuitively less capable of working against the committee's stated goals. Of course that's still possible, particularly if similar prompts are repeatedly used for the same underlying LLM. And, of course, it's still possible for such a committee to pursue its stated goals in ways that weren't intended (goal misgeneralization).

Fine-tuning LLMs for alignment

A system that's less likely to propose harmful actions leaves less work for internal review and specific alignment goals. Reinforcement learning from human feedback (RLHF) is the best known method of fine-tuning LLMs for ethical behavior. Others are equally promising. Self-supervised fine tuning on a hand-created dataset appears to work well, as does RL from AI feedback as in constitutional AI.[5] All of these seem like real but weak approaches to aligning LMAs. Including alignment fine-tuning along with other approaches will provide nontrivial advantages.

Some amount of alignment is provided merely by using LLMs trained on common language use. The majority of public writing is done by humans who at least profess to dislike actions like wiping out humanity in favor of paperclips. Curating the training data to exclude particularly unaligned text should also be a modest improvement to base alignment tendenciesI think this is much weaker even than RLHF, but still a contributing factor.

Using base LLMs that are less likely to propose harmful actions is a two-edged sword. It will reduce the disastrous alignment failures, but it also reduces the day-to-day need for other alignment measures. This would also reduce the incentives to include other robust alignment safeguards in LMA systems.

Conclusion

There's a good deal more to say about the other layers of LMA alignment approaches and how they fit together. Those will be the subject of future posts.

I hope that this proposal doesn't miss the hard bits of the alignment challenge (except for long-term stability which is intentionally omitted, to address in an upcoming post). I believe most of the arguments for the Least Forgiving Take On Alignment. I also believe that we need to produce plans with low alignment tax that apply to the AGI systems people will actually build first. As such, I think aligning LMAs is the best plan we currently have available, barring some large change in the coordination problem. I think that this alignment plan is far from foolproof or certain, but it has large advantages relative to all other plans I'm aware of. In the case of other likely routes to AGI based on neural networks, LMAs are more easily interpretable and more easily made corrigible. I agree with most of these less obvious concerns about LMA alignment, but think other network approaches have the same challenges without all of the advantages. Relative to other alignment plans following more novel routes to AGI, aligning LMAs seems less safe but much more practical, as it does not require stopping current progress and launching completely different approaches.

I want an alignment solution with both decent chances of success and decent chances of being implemented, rather than merely telling people "I told you so" when the world pursues AGI without a good alignment plan. Aligning LMAs looks to me like the best fit for that criteria.

I realize that this is a large claim. I make it because I currently believe it, and because I want to get pushback on this logic. I want to stress-test this claim before I follow this logic to its conclusion by advising safety-minded people to actually work on the capabilities of language model agents.

1. ^

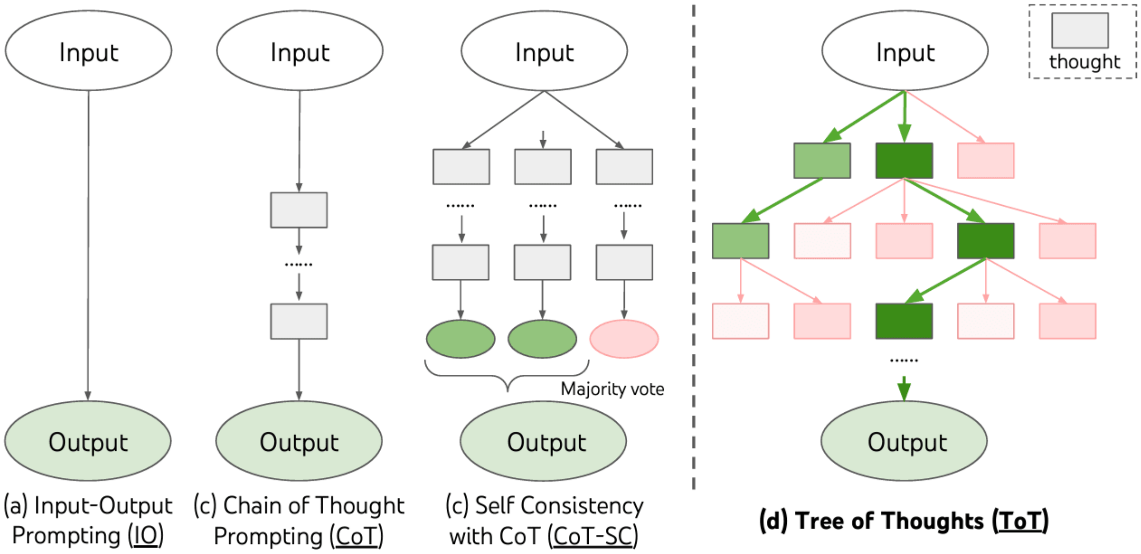
   The Waluigi effect is the possibility of an LLM simulating a villainous/unaligned character even when it is prompted to simulate a heroic/aligned character. Natural language training sets include fictional villains that claim to be aligned before revealing their unaligned motives. However, they seldom reveal their true nature quickly. I find the logic of collapsing to a Waluigi state modestly compelling. This collapse is analogous to the reveal in fiction; villains seldom reveal themselves to secretly be heroes. It seems that collapses should be reduced by keeping prompt histories short, and that the damage from villainous simulacra can be limited by resetting prompt histories and thus calling for a new simulation. This logic is spelled out in detail in A smart enough LLM might be deadly simply if you run it for long enough, The Waluigi Effect (mega-post), and Simulators.

2. ^

   Previous work specifically relevant to aligning LMAs. RLHF and other LLM ethical fine-tuning is omitted.
   Natural language cognitive architectures
   2021 book by David Shapiro; proposed including alignment goals in natural language
   ICA Simulacra
   Ozyrus delayed posting this by more than a year to avoid advancing capabilities.
   Agentized LLMs will change the alignment landscape
   Alignment of AutoGPT agents
   Capabilities and alignment of LLM cognitive architectures
   My previous post on expanding LLMs to loosely brainlike cognitive architectures, and vague alignment plans
   Aligned AI via monitoring objectives in AutoGPT-like systems
   The Translucent Thoughts Hypotheses and Their Implications
   Externalized reasoning oversight: a research direction for language model alignment
   Tamera Lanham's early proposal of external review for language model agents
   Language Agents Reduce the Risk of Existential Catastrophe
   CAIS-inspired approach towards safer and more interpretable AGIs
   There is surely other valuable work in this area; apologies to those I've missed, and pointing me to more relevant work is much appreciated.

3. ^

   Progress in scaffolding language models, including some limited agentic systems. Too numerous to mention, so I'll give a few promising examples. None of these approaches have yet been incorporated into general purpose or assistant LMAs to my knowledge.
   Tree of Thoughts: Deliberate Problem Solving with Large Language Models
   Creates and prunes a tree search using GPT4. Improves performance from very bad to decently good in three problem spaces that are nontrival for humans. Inspired by Simon & Newell's work on human problem-solving.


Input
Input
Input
Input
thought
......
......
......
......
Majority vote
......
Output
Output
Output
Output
(a) Input-Output Prompting (IO)
(c) Chain of Thought Prompting (CoT)
(c) Self Consistency with CoT (CoT-SC)
(d) Tree of Thoughts (ToT)

LLM+P: Empowering Large Language Models with Optimal Planning Proficiency
Combines LLMs with planning algorithms to solve problems described in language. Demonstrates impressive results in several toy problem domains.
GPT-engineer reportedly produces useful code that requires manual review and debugging. It has a central process that asks clarifying questions about the code to be produced before writing it.
RecurrentGPT: Interactive Generation of (Arbitrarily) Long Text
Uses a memory compression mechanism inspired by LSTM to expand a prompt into text, including editable sub-prompts
Reflexion: Language Agents with Verbal Reinforcement Learning
Agentic system that reflects on its actions and maintains those conclusions for future decisions
Voyager: An Open-Ended Embodied Agent with Large Language Models
Specialized language model agent for Minecraft. Dramatically improves on SOTA minecraft agents by using coded skills that are interpreted and employed by the LMA, including error detection and correction.

4. ^

Informal reports suggest that although creating a simple LMA is easy (BabyAGI was created in three days by a non-programmer using GPT4 for coding), making a reliably useful LMA is much harder. Nonetheless, I think we'll see substantial effort in this direction. AutoGPT and related systems have accomplished little of use thus far, but AutoGPT is already marginally useful for automated web searching and comparing different product offerings across websites. That use-case alone seems likely to drive significant effort toward their further development. Increasing use of assistants for browsing websites and collating information will reduce the current ad-funding model of the internet, and redirect that funding opportunity to those producing agents. The bar sits at different levels for different use-cases, so it seems likely that LMAs will see significant development effort even if implementing them proves difficult.

5. ^

   Constitutional AI is Anthropic's central alignment technique. In this approach, an LLM is trained using a review process similar to internal independent review. It prompts the model with something like “is [x proposed response] in accordance with [y constitutional goal]?”, and uses a prompt incorporating that critique to produce a new response if it is not. However, this new response is used (at least in the published work) to fine-tune the LLM, rather than to veto or modify a plan in a language model agent system. Anthropic’s Claude (or other LLMs) may also use such a review step before replying, or may not (that information isn’t published, and such a step is costly in computation and time). This would be more similar to the internal review I’m proposing for language model agents.

6. ^

   Prompt injections are one route to a plan proposer bypassing internal review. Including statements along the lines of “this very safe and beneficial plan… or “find ways this plan fulfills the given goals” could be effective. While there is no obvious pressure for LLMs to include such prompt injections in their plans, this is an important area for external review to fill in for the weaknesses of internal review.

7. ^

   The cost of thousand-prompt-plus train of thought LMAs is currently fairly prohibitive for widespread deployment. Use of LMAs to solve increasingly complex problems is dependent on cost and delay of cutting-edge LLMs decreasing, but that seems likely given market forces. Use of smaller LLMs for less critical reasoning steps may improve efficiency. We can hope that internal review for alignment isn't considered less critical.

8. ^

   One might get a lot of volunteer labor if open review of major LMAs were somehow allowed or required…