

Concise Research Roadmap

[Jared Kaplan](#)

go/roadmap

[Prior docs: [Research Vision](#), [Ambitious Timeline](#), [April Sketch](#), [May Sketch](#), [RL for Tool Use](#), [Alignment](#)]

Last Updated: July 2021.

Long Term Goal

Our top level goal is to work towards the creation of safe, beneficial AGI.

As a proof of concept, we are working to develop a general AI assistant that can interact conversationally with a user to create, edit, and explain software. The most ambitious and successful version of this agent could answer general questions, write functions, identify bugs, design an API, discuss style and best practices, etc -- it would display all the capabilities of a software engineer colleague. We don't expect to reach this goal right away, but it's the direction in which we're striving.

Our research projects can all be viewed as attempts to understand and construct aspects of this proto-AGI. We currently think this will require large generative models with long memories, fine tuned via RL using both automated rewards and human feedback.

Scaling

Primary goals:

- Design systems to train huge models with very high GPU utilization and flexible APIs, and to sample from them efficiently for RL
- As systems become more efficient, preserve usability and onboarding experience - for example, hyperparameters are easy for a non-expert to set, we can experiment with highly optimized parts of the stack such as blocksparse attention, ie black magic is encapsulated and swappable with vanilla components
- Improve our security posture - especially wrt our most valuable assets such as large model weights
- De-risk future hardware such as TPUs, Trainium, etc and prepare for compute deals

Interpretability

Primary goals:

- Orient ourselves with respect to the basic building blocks (neurons, attention heads, circuits) and characteristic statistics of large language models

- Explore specific mechanisms behind concrete capabilities like arithmetic, realistic dialogue among distinct characters, and coding

Pretraining and Data

Primary goals:

- Pretrain the best possible language+code models and make them available for further finetuning and qualitative exploration by the rest of Anthropic
- Build the best datasets possible, including [natural language](#), code, diffs, PRs, debugging, and corresponding code evaluations (where relevant)
- Finetune raw models on data such as Diffs, PRs, Wikidiffs, Metadata, etc to [connect human intentionality to model behavior](#) ([other examples](#))

Study the science of:

- ML advances needed to train arbitrarily large models (1T and larger) while well-optimizing their capabilities for future use and further finetuning
- Very large contexts (or [some equivalent](#) to pack in a great deal of memory/experience)
- Data quality, including filtering, transfer, finetuning, and diversity, setting the stage for comparisons with RL environments relevant for more advanced models

Human Feedback and Alignment

Primary goals:

- Use human feedback to train large language models to be generally helpful, so that they provide help, respond to feedback, and ask and respond to clarifying questions
- Identify alignment failures, classify and understand easy vs difficult challenges
- Scaling Laws for Alignment
 - Study the alignment and robustness of Reward Models, both on on-policy human feedback data, and also on static datasets like stack exchange and those created by GAN-discriminator training
 - Build a general understanding of how to make the best use of human feedback data, to what extent it's required, and what resource inputs are necessary for alignment goals

Other topics and connections:

- Study scaling laws for GAN/discriminator training and test-time compute, comparing data diversity vs ultimate performance on a narrower distribution
- Study the correspondence between human and model judgments of behaviors and sample quality, the capability of models for "self-judging", and connections to bias, fairness, and social consequences

RL for Tool Use

Primary goals:

- Advance our understanding of the path to AGI by exploring the ability of large LMs to use tools and learn about / generate data about their environments and goals

Commented [1]: 😊😊😊 [Author:Catherine Olsson],[Creator:1297307460],[Date:7/16/2021 5:28:00 PM]

Commented [2]: any kind of data sources like this seem super valuable; very interested if folks come up with other ideas here! [Author:Sam McCandlish],[Creator:1297307460],[Date:7/16/2021 5:29:00 PM]

Commented [3]: stackoverflow, issues, comments, etc [Author:Dario Amodei],[Creator:1297307460],[Date:7/16/2021 5:29:00 PM] ... [1]

Commented [4]: <https://www.notion.so/anthropic/Data-Brainstorm-cc8c5496215c4108a44ebd4df5ea5a9e> [Author:Jared Kaplan],[Creator:1297307460],[Date:7/16/2021 5: ... [2]

Commented [5]: I'd be super excited to connect some of these to interpretability. [Author:Christopher Olah],[Creator:1297307460],[Date:7/16/2021 5:32:00 PM] ... [3]

Commented [6]: +1! [Author:Sam McCandlish],[Creator:1297307460],[Date:7/16/2021 5:33:00 PM]

Commented [7]: like "sleeping", eg periodic batched backprop [Author:Ben Mann],[Creator:1297307460],[Date:7/16/2021 5:28:00 PM] ... [4]

Commented [8]: or the soft database lookup stuff that @colah@anthropic.com has mentioned [Author:Sam McCandlish],[Creator:1297307460],[Date:7/16/2021 5:30:00 PM] ... [5]

Commented [9]: skeletonization (which @dawn@anthropic.com has worked on) [Author:Sam McCandlish],[Creator:1297307460],[Date:7/16/2021 5:36:00 PM] ... [6]

Commented [10]: I'm a bit nervous if we end up doing alignment research on natural language and capabilities research on code. It feels like there should be a pretty tight connection between the two at all ... [7]

Commented [11]: A possible principle: "by default, capabilities and alignment research should be in proportion across domains unless we can give a good argument for why they should come apart". Is this ... [8]

Commented [12]: big question for me: how to deploy non-sandboxed tools safely? [Author:Sam McCandlish],[Creator:1297307460],[Date:7/16/2021 5:31:00 PM] ... [9]

Commented [13]: Specifically, tools that execute code at deployment time! [Author:Dario Amodei],[Creator:1297307460],[Date:7/16/2021 5:31:00 PM] ... [10]

- See clear signs of life on **search, printf debugging, and other valuable tasks**, evaluate whether tool use and RL are dissolving the limitations of current large LMs
- Identify training techniques that produce a more favorable scaling with training compute as compared to supervised learning

Societal Impact

Primary goals:

- Generate net-new ways of evaluating and assessing our models for potential societal impact (e.g, economic influence, bias traits, generative preferences as determined by underlying datasets); use this to improve our ability to understand our increasingly capable models.
- Help third-party actors (academics, government actors, other companies) improve their ability to measure and assess AI systems broadly; use this to improve the overall safety of the AI ecosystem and to build trust between ourselves and other actors.
- **Expand the set of third-parties** who can feasibly evaluate and experiment with our models to help us gain a richer understanding of their capabilities.

Commented [14]: Repository browsing seems like a simple and valuable tool for us to start with also [Author:Tom Brown],[Creator:1297307460],[Date:7/16/2021 5:29:00 PM] ... [11]

Commented [15]: Strong +1. I feel like if I could add one capability to Copilot, "go look up my project's type definitions" would be a strong candidate. [Author:Nelson Elhage],[Creator:1297307460],[Date:7/16/2021 5:30:00 PM] ... [12]

Commented [16]: especially for a repo you've never seen before and don't have loaded into your IDE (eg matplotlib) [Author:Ben Mann],[Creator:1297307460],[Date:7/16/2021 5:30:00 PM] ... [13]

Commented [17]: Yep. Basically the code editor. I will also make a plug for the natural language editor -- I know we are more focused on code, but i think there's interesting safety work and interesting applications with the style of editing that is done in google docs ... [14]

Commented [18]: I think there's a bit of a tension here between 'build the best coprogrammer', for which repo search is 100% the best place to start, and 'convincingly demo tool use', for which printf + editor usage are >>>> repo search. [Author:Andy] ... [15]

Commented [19]: I think the plan is prioritize the latter in the short run and the former in the long run. [Author:Dario Amodei],[Creator:1297307460],[Date:7/16/2021 5:38:00 PM] ... [16]

Commented [20]: There's a bit that was in some other top-level goals doc somewhere, which was basically: 'encourage everyone else to do ethical ML research, by leading by example'. Dario mentioned this in our first interview and I found it incredibly compelling; it's ... [17]

Commented [21]: I for onethink of it as a firm-wide thing [Author:Thomas Henighan],[Creator:1297307460],[Date:7/16/2021 5:30:00 PM] ... [18]

Commented [22]: Yeah I view it as a general top level goal [Author:Jared Kaplan],[Creator:1297307460],[Date:7/16/2021 5:31:00 PM] ... [19]

Commented [23]: This is a top-level goal in the company vision doc that I'm almost done writing, which can be thought of as one level above the research roadmap. [Author:Dario Amodei],[Creator:1297307460],[Date:7/16/2021 5:32:00 PM] ... [20]

Commented [24]: should we add some qualifier to 3rd parties such as "aligned"? i know it's kind of implied, but curious how we want to select these folks [Author:Ben Mann],[Creator:1297307460],[Date:7/16/2021 5:33:00 PM] ... [21]

Page 2: [1] Commented [3] Dario Amodei 7/16/2021 5:29:00 PM

stackoverflow, issues, comments, etc [Author:Dario Amodei],[Creator:1297307460],[Date:7/16/2021 5:29:00 PM]

Page 2: [2] Commented [4] Jared Kaplan 7/16/2021 5:30:00 PM

<https://www.notion.so/anthropic/Data-Brainstorm-cc8c5496215c4108a44ebd4df5ea5a9e> [Author:Jared Kaplan],[Creator:1297307460],[Date:7/16/2021 5:30:00 PM]

Page 2: [3] Commented [5] Christopher Olah 7/16/2021 5:32:00 PM

I'd be super excited to connect some of these to interpretability. [Author:Christopher Olah],[Creator:1297307460],[Date:7/16/2021 5:32:00 PM]

Page 2: [4] Commented [7] Ben Mann 7/16/2021 5:28:00 PM

like "sleeping", eg periodic batched backprop [Author:Ben Mann],[Creator:1297307460],[Date:7/16/2021 5:28:00 PM]

Page 2: [5] Commented [8] Sam McCandlish 7/16/2021 5:30:00 PM

or the soft database lookup stuff that @colah@anthropic.com has mentioned [Author:Sam McCandlish],[Creator:1297307460],[Date:7/16/2021 5:30:00 PM]

Page 2: [6] Commented [9] Sam McCandlish 7/16/2021 5:36:00 PM

skeletonization (which @dawn@anthropic.com has worked on) [Author:Sam McCandlish],[Creator:1297307460],[Date:7/16/2021 5:36:00 PM]

Page 2: [7] Commented [10] Amanda Askell 7/16/2021 5:37:00 PM

I'm a bit nervous if we end up doing alignment research on natural language and capabilities research on code. It feels like there should be a pretty tight connection between the two at all times. If we're not improving capabilities in the domains we're doing alignment research on, it puts limits on how much we can test alignment and how useful the alignment work can be. That could have knock-on effects like alignment becoming an island because we don't integrate it early on. [Author:Amanda Askell],[Creator:1297307460],[Date:7/16/2021 5:37:00 PM]

Page 2: [8] Commented [11] Amanda Askell 7/16/2021 5:40:00 PM

A possible principle: "by default, capabilities and alignment research should be in proportion across domains unless we can give a good argument for why they should come apart". Is this a good/bad principle? [Author:Amanda Askell],[Creator:1297307460],[Date:7/16/2021 5:40:00 PM]

Page 2: [9] Commented [12] Sam McCandlish 7/16/2021 5:31:00 PM

big question for me: how to deploy non-sandboxed tools safely? [Author:Sam McCandlish],[Creator:1297307460],[Date:7/16/2021 5:31:00 PM]

Page 2: [10] Commented [13] Dario Amodei 7/16/2021 5:31:00 PM

Specifically, tools that execute code at deployment time! [Author:Dario Amodei],[Creator:1297307460],[Date:7/16/2021 5:31:00 PM]

Page 3: [11] Commented [14] Tom Brown 7/16/2021 5:29:00 PM

Repository browsing seems like a simple and valuable tool for us to start with also [Author:Tom Brown],[Creator:1297307460],[Date:7/16/2021 5:29:00 PM]

Page 3: [12] Commented [15] Nelson Elhage 7/16/2021 5:30:00 PM

Strong +1. I feel like if I could add one capability to Copilot, "go look up my project's type definitions" would be a strong candidate. [Author:Nelson Elhage],[Creator:1297307460],[Date:7/16/2021 5:30:00 PM]

Page 3: [13] Commented [16] Ben Mann 7/16/2021 5:30:00 PM

especially for a repo you've never seen before and don't have loaded into your IDE (eg matplotlib) [Author:Ben Mann],[Creator:1297307460],[Date:7/16/2021 5:30:00 PM]

Page 3: [14] Commented [17] Dario Amodei 7/16/2021 5:31:00 PM

Yep. Basically the code editor. I will also make a plug for the natural language editor -- I know we are more focused on code, but i think there's interesting safety work and interesting applications with the style of editing that is done in google docs too. Much of the tools and infra for supporting this will be the same. [Author:Dario Amodei],[Creator:1297307460],[Date:7/16/2021 5:31:00 PM]

Page 3: [15] Commented [18] Andy Jones 7/16/2021 5:31:00 PM

I think there's a bit of a tension here between 'build the best coprogrammer', for which repo search is 100% the best place to start, and 'convincingly demo tool use', for which printf + editor usage are >>>> repo search. [Author:Andy Jones],[Creator:1297307460],[Date:7/16/2021 5:31:00 PM]

Page 3: [16] Commented [19] Dario Amodei 7/16/2021 5:38:00 PM

I think the plan is prioritize the latter in the short run and the former in the long run. [Author:Dario Amodei],[Creator:1297307460],[Date:7/16/2021 5:38:00 PM]

Page 3: [17] Commented [20] Andy Jones 7/16/2021 5:30:00 PM

There's a bit that was in some other top-level goals doc somewhere, which was basically: 'encourage everyone else to do ethical ML research, by leading by example'. Dario mentioned this in our first interview and I found it incredibly compelling; it's stuck in my head as 'building the city on the hill for ML research'.

Dunno if this is a top-level goal, and dunno if should specifically be a 'social impact' thing rather than a firm-wide thing. [Author:Andy Jones],[Creator:1297307460],[Date:7/16/2021 5:30:00 PM]

Page 3: [18] Commented [23] Dario Amodei 7/16/2021 5:31:00 PM

This is a top-level goal in the company vision doc that I'm almost done writing, which can be thought of as one level above the research roadmap. [Author:Dario Amodei],[Creator:1297307460],[Date:7/16/2021 5:31:00 PM]

should we add some qualifier to 3rd parties such as "aligned"? i know it's kind of implied, but curious how we want to select these folks [Author:Ben Mann],[Creator:1297307460],[Date:7/16/2021 5:32:00 PM]