### What we aim to do

Build software products which help people to save the world. We have various specific ideas which seem promising, but at a high level we want to spend time locating a major challenge to the ecosystem we can address with our skills, and iterate on a product to solve it.

We will focus most of our dev time on the project we think has the highest EV, and aim to dramatically improve some notable aspect of the ecosystem. Stampy has been our high EV bet so far, and is becoming [a single-point-of-access to AI Safety](https://docs.google.com/document/d/1KaEYEkUN7PEUmUZ_ilLV0OeNCdBhRyn6Ua8DYpwx1Jg/edit).

We will keep our feelers out for low-hanging fruit which can be solved in a few days with (now surprisingly good!) no-code apps. [Superlinear Prizes](https://www.super-linear.org/)[[1]](#footnote-0), [Effective Altruism Domains](https://ea.domains/), and [alignment.dev](https://alignment.dev/) are examples of micro projects our team has built already with <1 week combined dev time.

We will test out matchmaking volunteer developers with projects and incubating the outcomes. We think we can create a Schelling point where programmers who want to help and promising ideas can connect. If we’re right, we can bring useful tools into existence, without spreading ourselves thin by allocating our paid developers to too many projects.

### What we will build

We’re excited to create a unified personalized feed for AI safety to route the flood of information caused by the field’s rapid growth to the right people. Personalized feeds are the convergent solution to having too much content for any individual to read. We would pull from [every relevant source](https://github.com/moirage/alignment-research-dataset#sources), and optimize for intellectual growth when presenting content.

Existing alignment information sources (Rohin Shah’s newsletter, LessWrong.com, Alignment Forum) tend to surface mainly the most recent developments. We hypothesize that a better system would surface the best information for a user, given their prior knowledge of alignment and their preferred method of media consumption: video, article, post, paper, etc.

We can collect information about how a piece helps readers grow intellectually with carefully chosen ‘reactions’[[2]](#footnote-1) to let us optimize for creating value, not just time-on-site or other easily Goodharted metrics. Support for distillations and short-form content has been requested by potential users, as it would make it easier to stay up to date, but even better options for feature directions may emerge with more feedback. In parallel, we will have regular calls with community members to further clarify our models of how to develop the product.

Our current sketch of a roadmap looks like:

* Stage 1 - Database: Collect everything (using code from the [alignment dataset’s scrapers](https://github.com/moirage/alignment-research-dataset)) into a continuously updating and more comprehensive version of [QURI's AI Safety Papers database](https://ai-safety-papers.quantifieduncertainty.org/)
* Stage 2 - Basic feed: Stream of incoming info, core UI features, search (likely extending Stampy’s in-browser tensorflow semantic search)
* Stage 3 - Personalized feed: Collecting info from reacts and other user interactions to give them information which helps them grow intellectually, A/B test the default feed extensively
* Stage 4 (after having lots of user interviews to determine which direction to take the product) - Distillation features: Support for crowdsourced distillations and tl:drs

We think that this system could provide immense value by bringing each researcher and student information which is more specifically valuable to them. This would have a multiplicative effect on the overall information ecosystem, which will become increasingly significant as the field scales.

We have [a much longer but not super polished document](https://docs.google.com/document/d/1njNaCrhpRq4pZWddysMYZeuUHZBlin3g1OUNfaBGJ08/edit#) using the Lean Startup Canvas method to analyze the alignment ecosystem’s major challenges, which outlines some of our thought processes which lead us to this.

1. We built the backend and UI for the platform, with Nonlinear having the idea, writing text, and giving lots of input on the design. [↑](#footnote-ref-0)
2. Early ideas include up and down versions of ‘agree’, ‘well written’, ‘groked this’, ‘ambitious’, ‘on-topic’. Exact reactions and methods of feedback will be extensively discussed and tested. [↑](#footnote-ref-1)