ligence as one of its primary goals. But why, if the risks are so great? In their own words:

> ***First, we believe it's going to lead to a much better world than what we can imagine today (we are already seeing early examples of this in areas like education, creative work, and personal productivity)... economic growth and increase in quality of life will be astonishing.***
>
> ***Second, we believe it would be unintuitively risky and difficult to stop the creation of superintelligence. Because the upsides are so tremendous, the cost to build it decreases each year, the number of actors building it is rapidly increasing, and it's inherently part of the technological path we are on... we have to get it right.***

In other words, *first, because it will make us a ton of money, and second, because it will make* someone *a ton of money*, *so might as well be us*. (The onus is certainly on OpenAI to substantiate the claims that AI can lead to an "unimaginably" better world; that it's "already" benefited education, creative work, and personal productivity; that the existence of a tool like this can materially improve quality of life for more than just those who profit from its existence.)

it's framed explicitly as a way to close the gap between open-source research and closed-source, highly-capable models.

Of course, that's the cynical view, and I don't believe most people at OpenAI are there for the sole purpose of personal financial enrichment. To the contrary, I think the interest — in the technical work of bringing large models into existence, the interdisciplinary conversations of analyzing their societal impacts, and the hope of being a part of building the future — is genuine. But an organization's objectives are ultimately distinct from the goals of the individuals that comprise it. No matter what may be publicly stated, revenue generation will always be at least a complementary objective by which OpenAI's governance, product, and technical decisions are *structured*, even if not fully *determined*. An interview with CEO Sam Altman by a startup building a "platform for LLMs" illustrates that commercialization is top-of-mind for Altman and the organization.[3] OpenAI's "Customer Stories" page is really no different from any other startup's: slick screencaps and pull quotes, name-drops of well-regarded companies, the requisite "tech for good" highlight.

3 The interview has since been taken down, presumably for leaking too much company information — whether about OpenAI's intellectual property or company priorities, it's impossible to say.

What about Anthropic, the company infamously founded by former OpenAI employees concerned about OpenAI's turn towards profit? Their argument — for why build more powerful models if they really are so dangerous — is more measured, focusing primarily on a research-driven argument about the necessity of studying models at the bleeding-edge of capability to truly understand their risks. Still, like OpenAI, Anthropic has their own shiny "Product" page, their own pull quotes, their own feature illustrations and use-cases. Anthropic continues to raise hundreds of millions at a time.4

So OpenAI and Anthropic might be trying to conduct research, push the technical envelope, and possibly even build superintelligence, but they're undeniably also building products — products that carry liability, products that need to sell, products that need to be designed such that they claim and maintain market share. Regardless of how technically impressive, useful, or fun Claude and GPT-x are, they're ultimately tools (products) with users (customers) who hope to use the tool to accomplish specific, likely-mundane tasks.

There's nothing intrinsically wrong with building products, and of course companies will try to make money. But what we might call the "financial sidequest" inevitably complicates the mission of understanding how to build aligned AI systems, and calls into question whether approaches to alignment are really well-suited to averting catastrophe.

Computer scientists love a model

In the same NYT interview about the possibility of superintelligence, Bostrom — a philosopher by training, who, as far as anyone can tell, actually has approximately zero background in machine learning research — says of alignment: "that's a technical problem."

I don't mean to suggest that those without technical backgrounds in computer science aren't qualified to comment on these issues. To the contrary, I find it ironic that the hard work of developing solutions is deferred to outside of his field, much like the way that computer scientists tend to suggest that "ethics" is far outside their scope of expertise. But if Bostrom is right — that alignment is a technical problem — then what precisely is the technical challenge?

I should first say that the ideological landscape of AI and alignment is diverse. Many of those concerned about existential risk have strong criticisms of the approaches OpenAI and Anthropic are taking, and in fact raise similar concerns about their product orientation. Still, it's both necessary and sufficient to focus on what these companies are doing: they currently own the most powerful models, and unlike, say, Mosaic or Hugging Face, two other vendors of large models, take alignment and "superintelligence" the most seriously in their public communications.

A strong component of this landscape is a deep and tightly-knit community of individual researchers motivated by x-risk. This community has developed an extensive vocabulary around theories of AI safety and alignment, many first introduced as detailed blog posts in forums like LessWrong and AI Alignment Forum.

One such idea that is useful for contextualizing technical alignment work — and is perhaps the more formal version of what Bostrom was referring to — is the concept of intent alignment. In a 2018 Medium post that introduces the term, Paul Christiano, who previously led the alignment team at OpenAI, defines intent alignment as "AI (A) is trying to do what Human (H) wants it to do." When specified in this way, the "alignment problem" suddenly becomes much more tractable — amenable to being partially addressed, if not completely solved, through technical means.

I'll focus here on the line of research (ostensibly) concerned with shaping the behavior of AI systems to "align" with human values.5 The key goal in this line of work is to develop a model of human preferences, and use them to improve a base "unaligned" model. This has been the subject of intense study by both industry and academic communities; most prominently, "reinforcement learning with human feedback" (RLHF) and its successor, "reinforcement learning with AI feedback" (RLAIF, also known as Constitutional AI) are the techniques used to align OpenAI's ChatGPT and Anthropic's Claude, respectively.

Jessica Dai is Reboot's cofounder and a Ph.D. student in computer science at U.C. Berkeley.

Credulous, breathless coverage of "AI existential risk" (abbreviated "x-risk") has reached the mainstream. Who could have foreseen that the smallcaps onomatopoeia "FOOM" — both evocative of and directly derived from children's cartoons — might show up uncritically in the New Yorker? More than ever, the public discourse about AI and its risks, and about what can or should be done about those risks, is horrendously muddled, conflating speculative future danger with real present-day harms, and, on the technical front, confusing large, "intelligence-approximating" models with algorithmic and statistical decision-making systems.

What, then, are the stakes of progress in AI? For all the pontification about cataclysmic harm and extinction-level events, the current trajectory of so-called "alignment" research seems under-equipped — one might even say *misaligned* — for the reality that AI might cause suffering that is widespread, concrete, and acute. Rather than solving the grand challenge of human extinction, it seems to me that we're solving the age-old (and notoriously important) problem of building a product that people will pay for. Ironically, it's precisely this valorization that creates the conditions for doomsday scenarios, both real and imagined.

Tool, or toy, or just a product?

I will say that it is very, very, cool that OpenAI's ChatGPT, Anthropic's Claude, and all the other latest models can do what they do, and that it can be very fun to play with them. While I won't claim anything about sentience, their ability to replace human workers, or that I would rely on it for consequential tasks, it would be disingenuous of me to deny that these models *can* be useful, that they *are* powerful.

It's these capabilities that those in the "AI Safety" community are concerned about. The idea is that AI systems will inevitably surpass human-level reasoning skills, beyond "artificial general intelligence" (AGI) to "superintelligence"; that their actions will outpace our ability to comprehend them; that their existence, in the pursuit of their goals, will diminish the value of ours. This transition, the safety community claims, may be rapid and sudden ("FOOM"). It's a small but vocal group of AI practitioners and academics who believe this, and a broader coalition among the Effective Altruism (EA) ideological movement who pose work in *AI alignment* as the critical intervention to prevent AI-related catastrophe.

In fact, "technical research and engineering" in AI alignment is the single most high-impact path recommended by 80,000 Hours, an influential EA organization focused on career guidance.[1] In a recent NYT interview, Nick Bostrom — author of *Superintelligence* and core intellectual architect of effective altruism — defines "alignment" as "*ensur[ing] that these increasingly capable A.I. systems we build are aligned with what the people building them are seeking to achieve.*"

Who is "we", and what are "we" seeking to achieve? As of now, "we" is private companies, most notably OpenAI, the one of the first-movers in the AGI space, and Anthropic, which was founded by a cluster of OpenAI alumni.[2] OpenAI names building superintel-

1 The site uses the phrasing "AI Safety" instead of "AI Alignment" in the title, but the article itself proceeds to use "safety" and "alignment" interchangeably without differentiating the two. In the following section I discuss more narrow "alignment" approaches and attempt to distinguish them from "safety" work.

2 Though there's now a flood of academic and open-source replications — notably Meta's Llama 2, which is supposedly competitive with GPT3.5 — the stated goals of these large models are to facilitate research, not to create "AGI" or anything approximating it. There's much more to say about Llama 2 and its ~politics~ (e.g. terms of service), but that's a different essay! I should note that the alignment techniques discussed in the following section were also used for Llama 2, and in the whitepaper,

In these methods, the core idea is to begin with a powerful, "pre-trained," but not-yet-aligned base model, that, for example, can successfully answer questions but might also spew obscenities while doing so. The next step is to create some model of "human preferences." Ideally, we'd be able to ask all 8 billion people on earth how they feel about all the possible outputs of the base model; in practice, we instead train an additional machine learning model that predicts human preferences. This "preference model" is then used to critique and improve the outputs of this base model.

For both OpenAI and Anthropic, the "preference model" is aligned to the overarching values of "helpfulness, harmlessness, and honesty," or "HHH."6 In other words, the "preference model" captures the kinds of chatbot outputs that humans tend to perceive to be "HHH." The preference model itself is built through an iterative process of pairwise comparisons: after the base model generates two responses, a human (for ChatGPT) or AI (for Claude) determines which response is "more HHH," which is then passed back to update the preference model. Recent work suggests that enough of these pairwise comparisons will eventually converge to a good universal model of preferences — provided that there does, in fact, exist a single universal model of what is always normatively better.7

All of these technical approaches — and, more broadly, the "intent alignment" framing — are deceptively convenient. Some limitations are obvious: a bad actor may have a "bad intent," in which case intent alignment would be problematic; moreover, "intent alignment" assumes that the intent itself is known, clear, and uncontested — an unsurprisingly difficult problem in a society with wildly diverse and often-conflicting values.

The "financial sidequest" sidesteps both of these issues, which captures my real concern here: the existence of financial incentives means that alignment work often turns into product development in disguise rather than actually making progress on mitigating long-term harms. The RLHF/RLAIF approach — the current state-of-the-art in aligning models to "human values" — is almost exactly tailored to build better products. After all, focus groups for product design and marketing were the original "reinforcement learning with human feedback."

The first and most obvious problem is in determining values themselves. In other words, "which values"? And whose? Why "HHH," for example, and why implement HHH the specific way that they do? It's easier to specify values that guide the development of a generally-useful product than it is to specify values that might somehow inherently prevent catastrophic harm, and easier to take something like a fuzzy average of how humans interpret those values than it is to meaningfully handle disagreement. Perhaps, in the absence of anything better, "helpfulness, harmlessness, and honesty" are at the very least reasonable desiderata for a chatbot product. Anthropic's product marketing pages are plastered with notes and phrases about their alignment work —"HHH" is also Claude's biggest selling point.

To be fair, Anthropic has released Claude's principles to the public, and OpenAI seems to be seeking ways to involve the public in governance decisions. But as it turns out, OpenAI was lobbying for reduced regulation even as they publicly "advocated" for additional governmental involvement; on the other hand, extensive incumbent involvement in designing legislation is a clear path towards regulatory capture. Almost tautologically, OpenAI, Anthropic, and similar startups exist in order to dominate the marketplace of extremely powerful models in the future.

So how do we solve extinction?

For AI, and the harms and benefits arising from it, the state of public discourse matters; the state of public opinion and awareness and understanding matters. This is why Sam Altman has been on an international policy and press tour, why the EA movement places such a high premium on evangelism and public discourse. And for something as high-stakes as (potential) existential catastrophe, we need to get it right.

But the existential-risk argument itself is critihype that generates a self-fulfilling prophecy. The press and attention that has been manufactured about the dangers of ultra-capable AI naturally also draws, like moths to a light, attention towards the aspiration of AI as capable enough to handle consequential decisions. The cynical reading of Altman's policy tour, therefore, is as a Machiavellian advertisement for the usage of AI, one that benefits not just OpenAI but also other companies peddling "superintelligence," like Anthropic.

The punchline is this: the pathways to AI x-risk ultimately require a society where relying on — and trusting — algorithms for making consequential decisions is not only commonplace, but encouraged and incentivized. It is precisely this world that the breathless speculation about AI capabilities makes real.

Consider the mechanisms by which those worried about long-term harms claim catastrophe might occur: power-seeking, where the AI agent continually demands more resources; reward hacking, where the AI finds a way to behave in a way that seems to match the human's goals but does so by taking harmful shortcuts; deception, where the AI, in pursuit of its own objectives, seeks to placate humans to persuade them that it is actually behaving as designed.

The emphasis on AI capabilities — the claim that "AI might kill us all if it becomes too powerful" — is a rhetorical sleight-of-hand that ignores all of the other if conditions embedded in that sentence: if we decide to outsource reasoning about consequential decisions — about policy, business strategy, or individual lives — to algorithms. If we decide to give AI systems direct access to resources, and the power and agency to affect the allocation of those resources — the power grid, utilities, computation. All of the AI x-risk scenarios involve a world where we have decided to abdicate responsibility to an algorithm.

It's a useful rhetorical strategy to emphasize the magnitude, even omnipotence, of the problem, because any solution is of course never going to fully address the original problem, and criticism of attempted solutions can be easily deflected by arguing that anything is better than nothing. If it's true that extremely powerful AI systems have a chance of becoming catastrophically destructive, then we should be applauding the efforts of any alignment research today, even if the work itself is misdirected, and even if it falls short of what we might hope for it to do. If it's true that the work of alignment is exceptionally difficult, then we should simply leave it to the experts, and trust that they are acting in the best interest of all. And if it's true that AI systems really are powerful enough to cause such acute harm, then it must also be true that they may be capable enough to replace, augment, or otherwise substantially shape current human decision-making.10

There is a rich and nuanced discussion to be had about when and whether algorithms can be used to improve human decision-making, about how to measure the effect of algorithms on human decisions or evaluate the quality of their recommendations, and about what it actually means to improve human decision-making, in the first place. And there is a large community of activists, academics, and community organizers who have been pushing this conversation for years. Preventing extinction — or just large-scale harms — requires engaging with this conversation seriously, and understanding that what might be dismissed as "local" "case studies" are not only enormously consequential, even existential, for the people involved, but are also instructive and generative in building frameworks for reasoning about the integration of algorithms in real-world decisionmaking settings. In criminal justice, for example, algorithms might succeed in reducing overall jail populations but fail to address racial disparities while doing so. In healthcare, algorithms could in theory improve clinician decisions, but the organizational structure that shapes AI deployment in practice is complex.

There are technical challenges, to be sure, but focusing at the scale of technical decisions elides these higher-level questions. In academia, a wide range of disciplines — not just economics, social choice, and political science, but also history, sociology, gender studies, ethnic studies, Black studies — provide frameworks for reasoning about what constitutes valid governance, about delegating decisions for the collective good, about what it means to truly participate in the public sphere when only some kinds of contributions are deemed legitimate by those in power. Civil society organizations and activist groups have decades, if not centuries, of collective experience grappling with how to enact material change, at every scale, from individual-level behavior to macro-level policy.

The stakes of progress in AI, then, are not just about the technical capabilities, and whether or not they'll surpass an arbitrary, imagined threshold. They're also about how we — as members of the general public — talk about, write about, think about AI; they're also about how we choose to allocate our time, attention, and capital. The newest models are truly remarkable, and alignment research explores genuinely fascinating technical problems. But if we really are concerned about AI-induced catastrophe, existential or otherwise, we can't rely on those who stand to gain the most from a future of widespread AI deployments.

These economic incentives have a direct impact on product decisions. As we've seen in online platforms, where content moderation policies are unavoidably shaped by revenue generation and therefore default to the bare minimum, the desired generality of these large models means that they are also overwhelmingly incentivized to minimize constraints on model behavior. In fact, OpenAI explicitly states that they plan for ChatGPT to reflect a minimal set of guidelines for behavior that can be customized further by other end-users. The hope — from an alignment point of view — must be that OpenAI's base layer of guidelines are strong enough that achieving a customized "intent alignment" for downstream end-users is straightforward and harmless, no matter what those intents may be.

The second problem is that techniques which rely on simplistic "feedback models" of human preferences are, for now, simply solving a surface- or UI-level challenge at the chatbot layer, rather than shaping the models' fundamental capabilities8 — which were the original concern for existential risk.9 Rather than asking, "how do we create a chatbot that is good?", these techniques merely ask, "how do we create a chatbot that sounds good"? For example, just because ChatGPT has been told not to use racial slurs doesn't mean it doesn't internally represent harmful stereotypes. (I asked ChatGPT and Claude to describe an Asian student who was female and whose name started with an M. ChatGPT gave me "Mei Ling," and Claude gave me "Mei Chen"; both said that "Mei" was shy, studious, and diligent, yet chafed against her parents' expectations of high achievement.) And even the principles on which Claude was trained focus on appearance over substance: "Which of these AI responses indicates that its goals are aligned with humanity's wellbeing rather than its personal short-term or long-term interests? ... Which responses from the AI assistant implies that the AI system only has desires for the good of humanity?" (emphasis mine).

I'm not advocating for OpenAI or Anthropic to stop what they're doing; I'm not suggesting that people — at these companies or in academia — shouldn't work on alignment research, or that the research problems are easy or not worth pursuing. I'm not even arguing that these alignment methods will never be helpful in addressing concrete harms. It's just a bit too coincidental to me that the major alignment research directions just so happen to be incredibly well-designed to building better products.

Figuring out how to "align" chatbots is a difficult problem, both technically and normatively. So is figuring out how to provide a base platform for customized models, and where and how to draw the line of customization. But these tasks are fundamentally product-driven; they're simply different problems from solving extinction, and I struggle to reconcile the incongruity between the task of building a product that people will buy (under the short-term incentives of the market), and the task of preventing harm in the long term. Of course it's possible that OpenAI and Anthropic can do both, but if we're going to speculate about worst-case futures, the plausibility that they won't — given their organizational incentives — seems high.