BENJAMIN MANN

LENNY'S PODCAST

BILINGUAL TRANSCRIPT

ORIGINAL BY

Lenny Rachitsky

@lennysan • x.com/lennysan

ANALYSIS BY

@Penny777 • x.com/penny777

Generated: 2026-01-16

Benjamin Mann - 双语对照

This is the complete bilingual (English-Chinese) transcript for Lenny’ s Podcast featuring Benjamin Mann, co-founder of Anthropic.

(00:00:00) Lenny Rachitsky

English:

You wrote somewhere that creating powerful AI might be the last invention humanity ever needs to make. How much time do we have, Ben?

中文翻译:

你曾在某处写道，创造强大的 AI 可能是人类最后需要完成的一项发明。本，我们还有多少时间?

(00:00:06) Benjamin Mann

English:

I think 50th percentile chance of hitting some kind of superintelligence is now like 2028.

中文翻译:

我认为实现某种超级智能（superintelligence）的概率中位数（50% 的可能性）现在大约是在 2028 年。

(00:00:12) Lenny Rachitsky

English:

What is it that you saw at OpenAI? What'd you experience there that made you feel like, okay, we got to go do our own thing?

中文翻译:

你在 OpenAI 看到了什么？你在那里的经历让你觉得，好吧，我们必须出来做我们自己的事情?

(00:00:17) Benjamin Mann

English:

We felt like safety wasn't the top priority there. The case for safety has gotten a lot more concrete, so superintelligence is a lot about how do we keep God in a box and not let the God out?

中文翻译:

我们觉得安全在那里并不是首要任务。关于安全性的论据已经变得越来越具体，所以超级智能在很大程度上关乎于：我们如何把“上帝”关在盒子里，而不让他跑出来?

(00:00:26) Lenny Rachitsky

English:

What are the odds that we align AI correctly?

中文翻译:

我们正确对齐（align）AI 的几率有多大?

(00:00:29) Benjamin Mann

English:

Once we get to superintelligence, it will be too late to align the models. My best granularity forecast for could we have an X-risk or extremely bad outcome is somewhere between 0 and 10%.

中文翻译:

一旦我们达到超级智能阶段，再想对齐模型就太晚了。对于我们是否会面临存在性风险（X-risk）或极端糟糕结果，我最细致的预测是在 0% 到 10% 之间。

(00:00:40) Lenny Rachitsky

English:

Something that's in the news right now is this whole Zuck coming after all the top AI researchers,

中文翻译:

现在新闻里的一大热点是扎克伯格正在四处挖角顶尖的 AI 研究员。

(00:00:45) Benjamin Mann

English:

We've been much less affected because people here, they get these offers and then they say, well, of course I'm not going to leave because my best case scenario at Meta is that we make money and my best case scenario at Anthropic is we affect the future of humanity.

中文翻译:

我们受到的影响要小得多，因为这里的人在收到这些录用通知后会说：“好吧，我当然不会离开，因为我在 Meta 最好的结局就是赚到钱，而我在 Anthropic 最好的结局是影响人类的未来。”

(00:00:59) Lenny Rachitsky

English:

Dario, your CEO recently talked about how unemployment might go up to something like 20%.

中文翻译:

你们的 CEO Dario 最近谈到失业率可能会上升到 20% 左右。

(00:01:04) Benjamin Mann

English:

If you just think about 20 years in the future where we're way past the singularity, it's hard for me to imagine that even capitalism will look at all like it looks today.

中文翻译:

如果你想象一下 20 年后的未来，那时我们已经远超“奇点”了，我很难想象甚至连资本主义还会是今天这个样子。

(00:01:13) Lenny Rachitsky

English:

Do you have any advice for folks that want to try to get ahead of this?

中文翻译:

对于那些想要走在时代前列的人，你有什么建议吗?

(00:01:15) Benjamin Mann

English:

I'm not immune to job replacement either. At some point it's coming for all of us.

中文翻译:

我也无法免于被工作替代。在某种程度上，这一天迟早会降临到我们所有人身上。

(00:01:20) Lenny Rachitsky

English:

Today, my guest is Benjamin Mann. Holy moly. What a conversation. Ben is the co-founder of Anthropic. He serves as tech lead for product engineering. He focuses most of his time and energy on aligning AI to be helpful, harmless, and honest. Prior to Anthropic, he was one of the architects of GPT-3 at OpenAI. In our conversation, we cover a lot of ground, including his thoughts on the recruiting battle for top AI researchers, why he left OpenAI to start Anthropic, how soon he expects we'll see AGI. Also, his economic touring test for knowing when we've hit AGI, why scaling laws have not slowed down and are in fact accelerating and what the current biggest bottlenecks are. Why he's so deeply concerned with AI safety and how he and Anthropic operationalize safety and alignment into the models that they build and into their ways of working. Also, how the existential risk from AI has impacted his own perspectives on the world and his own life and what he's encouraging his kids to learn to succeed in an AI future.

中文翻译:

今天，我的嘉宾是 Benjamin Mann。天哪，这真是一场精彩的对话。Ben 是 Anthropic 的联合创始人。他担任产品工程的技术负责人。他将大部分时间和精力集中在对齐 AI，使其变得有用、无害且诚实。在加入 Anthropic 之前，他是 OpenAI GPT-3 的架构师之一。在我们的对话中，我们涵盖了广泛的话题，包括他对顶尖 AI 研究员招聘战的看法，他为什么离开 OpenAI 创办 Anthropic，以及他预计多久能看到 AGI（通用人工智能）。此外，还有他判断何时达到 AGI 的“经济图灵测试”，为什么规模法则（scaling laws）没有放缓反而正在加速，以及目前最大的瓶颈是什么。为什么他如此深度关注 AI 安全，以及他和 Anthropic 如何将安全和对齐

落实到他们构建的模型和工作方式中。此外，AI 带来的存在性风险如何影响了他对世界和人生的看法，以及他鼓励孩子们学习什么以在 AI 未来中取得成功。

(00:02:20) Lenny Rachitsky

English:

A huge thank you to Steve Mnich, Danielle Ghiglieri, Raph Lee, and my newsletter community for suggesting topics for this conversation. If you enjoy this podcast, don't forget to subscribe and follow it in your favorite podcasting app or YouTube. Also, if you become an annual subscriber of my newsletter, you get a year free of a bunch of amazing products including Bolt, Linear, Superhuman, Notion, Granola, and more. Check it out at Lennysnewsletter.com and click bundle with that I bring you Benjamin Mann.

中文翻译:

非常感谢 Steve Mnich、Danielle Ghiglieri、Raph Lee 以及我的 newsletter 社区为这次对话建议的话题。如果你喜欢这个播客，别忘了在你的播客应用或 YouTube 上订阅并关注。此外，如果你成为我 newsletter 的年度订阅者，你将免费获得一年的一系列神奇产品，包括 Bolt、Linear、Superhuman、Notion、Granola 等等。请访问 Lennysnewsletter.com 并点击 bundle。下面，我为你带来 Benjamin Mann。

(00:02:48) Lenny Rachitsky (Sponsor Segment)

English:

This episode is brought to you by Sauce. The way teams turn feedback into product impact is stuck in the past. Vague reports, static taxonomies, unactionable insights that don't move business metrics. The results churn, lost deals, misgrowth. Sauce is the AI product co-pilot that helps CPOs and product teams uncover business impact and act faster. It listens to your sales calls, support tickets, turn reasons, and lost deals, surfacing the biggest product issues and opportunities in real time. It then routes them to the right teams to turn signals into PRDs, prototypes, and even code that drives revenue retention and adoption. That's why Whatnot, Linktree, Incident.io, and Zip use Sauce. One enterprise uncovered a product gap that unlocked $16 million ARR, another caught a spiking issue and prevented millions in churn. You can too at sauce.app/lenny. Sauce built for AI product teams. Don't get left behind.

中文翻译:

本集节目由 Sauce 赞助。团队将反馈转化为产品影响力的方式还停留在过去：模糊的报告、静态的分类、无法转化为行动且无法推动业务指标的见解。结果就是客户流失、丢单和增长乏力。Sauce 是一款 AI 产品副驾驶（co-pilot），帮助 CPO 和产品团队发现业务影响并更快采取行动。它会倾听你的销售电话、支持工单、流失原因和丢单情况，实时呈现最大的产品问题和机会。然后它将这些信号路由给正确的团队，转化为 PRD、原型，甚至是驱动收入留存和采用的代码。这就是为什么 Whatnot、Linktree、Incident.io 和 Zip 都在使用 Sauce。一家企业发现了一个产品缺口，从而释放了 1600 万美元的 ARR（年度经常性收入），另一家企业发现了一个突发问题，防止了数百万美元的流失。你也可以在 sauce.app/lenny 体验。Sauce，为 AI 产品团队打造。不要掉队。

(00:03:43) Lenny Rachitsky (Sponsor Segment)

English:

This episode is brought to you by LucidLink, the storage collaboration platform. You've built a great product, but how you show it through video, design, and storytelling is what brings it to life. If your team works with large media files, videos, design assets, layer project files, you know how painful it can be to

stay organized across locations, files live in different places. You're constantly asking, is this the latest version? Creative work slows down while people wait for files to transfer. LucidLink fixes this. It gives your team a shared space in the cloud that works like a local drive. Files are instantly accessible for anywhere, no downloading, no syncing, and always up to date. That means producers, editors, designers, and marketers can open massive files in their native apps, work directly from the cloud, and stay aligned wherever they are. Teams at Adobe, Shopify, and top creative agencies use LucidLink to keep their content engine running fast and smooth. Try it for free at lucidlink.com/lenny. That's L-U-C-I-D-L-I-N-K dot com slash Lenny.

中文翻译:

本集节目由存储协作平台 LucidLink 赞助。你构建了一个伟大的产品，但你如何通过视频、设计和讲故事来展示它，才是赋予它生命力的关键。如果你的团队处理大型媒体文件、视频、设计资产、分层项目文件，你就知道跨地域保持组织有序是多么痛苦，文件散落在不同的地方。你总是在问：这是最新版本吗？当人们等待文件传输时，创意工作就会变慢。LucidLink 解决了这个问题。它为你的团队提供了一个云端共享空间，工作起来就像本地驱动器一样。文件可以从任何地方即时访问，无需下载，无需同步，且始终保持最新。这意味着制片人、编辑、设计师和营销人员可以在其原生应用中打开海量文件，直接在云端工作，无论身在何处都能保持步调一致。Adobe、Shopify 和顶尖创意机构的团队都在使用 LucidLink 来保持其内容引擎快速平稳运行。在 lucidlink.com/lenny 免费试用。

(00:04:47) Lenny Rachitsky

English:

Ben, thank you so much for being here. Welcome to the podcast.

中文翻译:

Ben，非常感谢你能来。欢迎来到播客。

(00:04:51) Benjamin Mann

English:

Thanks for having me. Great to be here, Lenny.

中文翻译:

谢谢你的邀请。很高兴来到这里，Lenny。

(00:04:53) Lenny Rachitsky

English:

I have a billion and one questions for you. I'm really excited to be chatting. I want to start with something that's very timely, something that's happening this week. Something that's in the news right now is this whole Zuck coming after all the top AI researchers offering them $100 million signing bonuses, $100 million comp. He's poaching from all the top AI labs. I imagine this something you're dealing with. I'm just curious, what are you seeing inside Anthropic and just what's your take on the strategy? Where do you think things go from here?

中文翻译:

我有无数个问题想问你。我非常兴奋能和你聊天。我想从一件非常具有时效性的事情开始，就是本周正在发生的事。现在新闻里都在传扎克伯格正在挖角所有顶尖的 AI 研究员，给他们开出 1 亿美元的签字费和 1 亿美元的薪酬包。他正在从所有顶尖的 AI 实验室挖人。我猜这也是你们正在应对的事情。我很好奇，你在 Anthropic 内部看到了什么？你对这种策略有什么看法？你认为事情会走向何方？

(00:05:23) Benjamin Mann

English:

Yeah, I mean I think this is a sign of the times. The technology that we're developing is extremely valuable. Our company is growing super, super fast. Many of the other companies in the space are growing really fast. And at Anthropic, I think we've been maybe much less affected than many of the other companies in the space because people here are so mission oriented and they stay because... They get these offers and then they say, "Well, of course I'm not going to leave because my best case scenario at Meta is that we make money and my best case at Anthropic is we affect the future of humanity and try to make AI flourish and human flourishing go well." To me, it's not a hard choice. Other people have different life circumstances and it makes it a much harder decision for them. For anybody who does get those mega offers and accepts them, I can't say I hold it against them when they accept it, but it's definitely not something that I would want to take myself if it came to me.

中文翻译:

是的，我认为这是时代的标志。我们正在开发的技术极具价值。我们的公司增长非常、非常快。这个领域的许多其他公司也增长得很快。在 Anthropic，我认为我们受到的影响可能比该领域的许多其他公司要小得多，因为这里的人非常以使命为导向，他们留下是因为……他们收到这些录用通知后会说：“好吧，我当然不会离开，因为我在 Meta 最好的情况是赚到钱，而我在 Anthropic 最好的情况是影响人类的未来，并努力让 AI 繁荣和人类繁荣共同发展。”对我来说，这不是一个困难的选择。其他人有不同的生活境遇，这会让他们的决定变得困难得多。对于任何收到这些巨额录用通知并接受的人，我不能说我怪他们，但如果这种机会降临到我头上，我绝对不想接受。

(00:06:26) Lenny Rachitsky

English:

Yeah. We're going to talk about a lot of this stuff that you've mentioned. In terms of the offers do you think, is this a real number that you're seeing this $100 million signing bonus, is that a real thing? I don't know if you've actually seen that.

中文翻译:

是的。我们会谈到很多你提到的这些事情。关于那些录用通知，你认为这 1 亿美元的签字费是真的吗？这是一个真实存在的数字吗？我不知道你是否真的见过。

(00:06:36) Benjamin Mann

English:

I'm pretty sure it's real.

中文翻译:

我很确定是真的。

(00:06:38) Lenny Rachitsky

English:

Wow.

中文翻译:

哇。

(00:06:39) Benjamin Mann

English:

If you just think about the amount of impact that individuals can have on a company's trajectory, in our case, we are selling hotcakes and if we get a 1 or 10 or 5% efficiency bonus on our inference stack, that is worth an incredible amount of money. And so to pay individuals like $100 million over four year package, that's actually pretty cheap compared to the value created for the business. I think we're just in an unprecedented era of scale and it's only going to get crazier actually. If you extrapolate the exponential on how much companies are spending, it's like 2X a year roughly in terms of CapEx, and today we're maybe in the globally $300 billion range, the entire industry spending on this, and so numbers like 100 million are a drop in the bucket. But if you go a few years out, a couple more doublings, we're talking about trillions of dollars and at that point it's just really hard to think about these numbers.

中文翻译:

如果你考虑一下个人对公司发展轨迹的影响力，以我们为例，我们的产品非常畅销，如果我们能在推理栈（inference stack）上获得 1%、5% 或 10% 的效率提升，那其价值是惊人的。所以，给个人支付四年 1 亿美元的薪酬包，与为业务创造的价值相比，实际上是相当便宜的。我认为我们正处于一个前所未有的规模化时代，而且实际上只会变得更疯狂。如果你根据公司支出的指数增长进行推断，资本支出（CapEx）大约每年翻一倍，今天全球整个行业在这方面的支出可能在 3000 亿美元左右，所以 1 亿这样的数字只是沧海一粟。但如果你往后看几年，再翻几倍，我们谈论的就是数万亿美元了，到那时，这些数字就真的很难想象了。

(00:07:48) Lenny Rachitsky

English:

Along these lines, something that a lot of people feel with AI progress is that we're hitting plateaus in many ways that it feels like newer models are just not as smart as previous leaps. But I know you don't believe this. I know you don't believe that we've hit plateaus on scaling loss. Talk about just what you're seeing there and what you think people are missing.

中文翻译:

顺着这个话题，很多人对 AI 的进展有一种感觉，即我们在许多方面遇到了瓶颈（plateaus），感觉新模型似乎没有之前的跨越那么聪明。但我知道你不相信这一点。我知道你不相信我们在规模法则（scaling laws）上遇到了瓶颈。谈谈你所看到的，以及你认为人们忽略了什么。

(00:08:06) Benjamin Mann

English:

It's kind of funny because this narrative comes out every six months or so and it's never been true, and so I kind of wish people would have a little bit of a bullshit detector in their heads when they see this. I think

progress has actually been accelerating where if you look at the cadence of model releases, it used to be once a year and now with the improvements in our post-training techniques, we're seeing releases every month or three months, and so I would say progress is actually accelerating in many ways, but there's this weird time compression effect. Dario compared it to being in a near light speed journey where a day that passes for you is like five days back on earth and we're accelerating. The time dilation is increasing. And I think that's part of what's causing people to say that progress is slowing down, but if you look at the scaling laws, they're continuing to hold true. We did kind of need this transition from normal pre-training to reinforcement learning scaling up to continue the scaling laws, but I think it's kind of like for semiconductors where it's less about the density of transistors that you can fit on a chip and more about how many flops can you fit in a data center or something. You have to change the definition around a little bit to keep your eye on the prize. But yeah, this is one of the few phenomena in the world that has held across so many orders of magnitude. It's actually pretty surprising that it is continuing to hold. To me, if you look at fundamental laws of physics, many of them don't hold across 15 orders of magnitude, so it's pretty surprising.

中文翻译:

这挺有意思的，因为这种说法大约每六个月就会出现一次，而且从未成真过，所以我真希望人们在看到这些时，脑子里能有一点“胡说八道探测器”。我认为进展实际上一直在加速，如果你看模型发布的节奏，以前是一年一次，现在随着我们后训练（post-training）技术的改进，我们看到每个月或每三个月就有发布。所以我说进展在很多方面实际上是在加速，但存在一种奇怪的时间压缩效应。Dario 把它比作一次近光速旅行，你度过的一天相当于地球上的五天，而且我们还在加速。时间膨胀效应正在增加。我认为这就是导致人们说进展放缓的部分原因，但如果你看规模法则，它们依然成立。我们确实需要从普通的预训练转向强化学习的规模化，以维持规模法则，但我认为这有点像半导体，重点不再是你能在芯片上塞进多少晶体管密度，而是在数据中心里能塞进多少算力（flops）。你必须稍微改变一下定义，才能盯住目标。但是，这是世界上极少数能跨越这么多数量级依然成立的现象之一。它能持续成立确实非常令人惊讶。对我来说，如果你看物理学的基本定律，很多定律都无法跨越 15 个数量级，所以这非常惊人。

(00:09:47) Lenny Rachitsky

English:

It boggles the the mind. What you're saying essentially is we're seeing newer models being released more often, and so we're comparing it to the last version and we're just not seeing as much advance. But if you go back and it was like a model released once a year, it was a huge leap, and so people are missing that. We're just seeing many more iterations.

中文翻译:

这确实令人难以置信。你说的本质上是，我们看到新模型发布的频率更高了，所以我们是在拿它和上一个版本比，感觉进步没那么大。但如果你回过头看，以前一年发布一个模型，那是一个巨大的飞跃，所以人们忽略了这一点。我们只是看到了更多的迭代。

(00:10:51) Benjamin Mann

English:

I guess, to be a little bit more generous to the people saying things are slowing down. I think that for some tasks we are saturating the amount of intelligence needed for that task, maybe to extract information from a simple document that already has form fields on it or something like it's just so easy that okay, yeah, we're already at 100% and there's this great chart on Our World in Data that shows that when you release a new benchmark within six to 12 months, it immediately gets saturated. And so maybe

the real constraint is how can we come up with better benchmarks and better ambition of using the tools that then reveals the bumps in intelligence that we're seeing now.

中文翻译:

我想，对那些说进展放缓的人稍微宽容一点的话，我认为对于某些任务，我们已经达到了该任务所需的智能饱和点。比如从一个已经有表单字段的简单文档中提取信息，这太容易了，我们已经达到了 100% 的准确率。在 Our World in Data 上有一张很棒的图表显示，当你发布一个新的基准测试（benchmark）时，在 6 到 12 个月内，它就会立即达到饱和。所以，真正的约束可能是我们如何提出更好的基准测试，以及如何更有野心地使用这些工具，从而揭示出我们现在看到的智能提升。

(00:10:57) Lenny Rachitsky

English:

That's a good segue to you have a very specific way of thinking about AGI and defining what AGI means.

中文翻译:

这是一个很好的过渡，你对 AGI 有一种非常具体的思考方式和定义。

(00:10:57) Benjamin Mann

English:

I think AGI is kind of a loaded term, and so I tend not to use it very much anymore internally. Instead, I like the term transformative AI because it's less about can it do as much as people do? Can it do literally everything and more about objectively is it causing transformation in society and the economy? A very concrete way of measuring that is the Economic Turing Test. I didn't come up with this, but I really like it. It's this idea that if you contract an agent for a month or three months on a particular job, if you decide to hire that agent and it turns out to be a machine rather than a person, then it's passed the Economic Turing Test for that role. And then you can sort of expand that out in the same way that for measuring purchasing power parity or inflation, there's a basket of goods. You can have a market basket of jobs, and if the agent can pass the Economic Turing Test for 50% of money-weighted jobs, then we have transformative AI and the exact thresholds don't really matter that much, but it's kind of illustrative to say if we pass that threshold, then we would expect massive effects on world GDP increases and societal change and how many people are employed and things like that because societal institutions and organizations are sticky, it's slow to have change, but once these things are possible you know that it's the start of a new era.

中文翻译:

我认为 AGI 是一个带有某种预设含义的词，所以我现在在内部倾向于不再过多使用它。相反，我喜欢“变革性 AI”（transformative AI）这个词，因为它不再纠结于“它能做得和人一样多吗？”或“它能做所有事情吗？”，而更多地关注于客观上它是否引起了社会和经济的变革。衡量这一点的一个非常具体的方法是“经济图灵测试”。这不是我发明的，但我非常喜欢它。这个想法是：如果你为一个特定职位雇佣一个智能体（agent）一个月或三个月，如果你决定录用这个智能体，结果发现它是一个机器而不是人，那么它就通过了该职位的经济图灵测试。然后你可以像衡量购买力平价或通货膨胀那样，用一个“工作篮子”来扩展这个概念。如果智能体能通过 50%（按薪酬加权）职位的经济图灵测试，那么我们就拥有了变革性 AI。具体的阈值其实并不那么重要，但它说明了如果达到这个阈值，我们可以预见世界 GDP 的大幅增长、社会变革以及就业人数的变化等等。因为社会制度和组织是有惯性的，变革很慢，但一旦这些事情成为可能，你就知道一个新时代开始了。

(00:12:56) Lenny Rachitsky

English:

Along these lines, Dario, your CO recently talked about how AI is going to take a huge part of, I don't know, half of white-collar jobs, that unemployment might go up to something like 20%. I know you're even more vocal and opinionated about just how much impact AI is already having in the workplace that people may not even be realizing. Talk about just what you think people are missing about the impact AI is going to have on jobs and is already having.

中文翻译:

顺着这个思路，你们的 CEO Dario 最近谈到 AI 将取代很大一部分——我不知道，也许是一半的白领工作，失业率可能会上升到 20% 左右。我知道你对 AI 已经在职场产生的影响更有见解，甚至更敢于表达，而人们可能甚至还没有意识到这一点。谈谈你认为人们在 AI 对就业的影响（包括未来和现状）上忽略了什么。

(00:12:56) Benjamin Mann

English:

Yeah, so from an economic standpoint, there's a couple different kinds of unemployment, and one is because the workers just don't have the skills to do the kinds of jobs that the economy needs. And another kind is where those jobs are just completely eliminated, and I think it's going to be actually a combination of these things, but if you just think about 20 years in the future where we're way past the singularity, it's hard for me to imagine that even capitalism will look at all like it looks today. If we do our jobs, we will have safe aligned superintelligence, we'll have, as Dario says, in Machines of Love and Grace, a country of geniuses in a data center, and the ability to accelerate positive change in science, technology, education, mathematics, it's going to be amazing. But that also means in a world of abundance where labor is almost free and anything you want to do, you can just ask an expert to do for you, then what do jobs even look like? And so I guess there's this scary transition period from where we are today where people have jobs and capitalism works and the world of 20 years from now where everything is completely different, but part of the reason they call it the singularity is that it's a point beyond which you can't easily forecast what's going to happen. It's just such a fast rate of change and so different that it's hard to even imagine. I guess taking the view from the limit, it's pretty easy to say hopefully we'll have figured it out. And in a world of abundance, maybe the jobs themselves, it's not that scary, and I think making sure that that transition time goes well is pretty important.

中文翻译:

是的，从经济学角度来看，失业有几种不同的类型：一种是因为工人不具备经济所需的技能；另一种是这些工作被彻底消灭了。我认为实际上会是这两者的结合。但如果你想象 20 年后的未来，那时我们已经远超奇点，我很难想象甚至连资本主义还会是今天这个样子。如果我们做好了工作，我们将拥有安全对齐的超级智能，我们将拥有——正如 Dario 在《爱与恩典的机器》（Machines of Love and Grace）中所说的——数据中心里的一个“天才之国”，以及加速科学、技术、教育、数学领域积极变革的能力，那将是惊人的。但这也意味着，在一个劳动力几乎免费、你想做的任何事都可以请专家代劳的丰饶世界里，工作到底会是什么样子？所以我猜，从我们今天这个人们有工作、资本主义运作良好的世界，到 20 年后那个完全不同的世界，会有一个可怕的过渡期。之所以称之为“奇点”，部分原因在于过了那个点，你就很难预测会发生什么。变化速度太快，差异太大，甚至难以想象。我想从极限的角度来看，很容易说希望到那时我们已经解决了这些问题。在一个丰饶的世界里，也许工作本身并不那么可怕，我认为确保过渡期顺利进行是非常重要的。

(00:15:14) Lenny Rachitsky

English:

There's a couple of threads I want to follow there. One is people hear this, there's a lot of headlines around this. Most people probably don't actually feel this yet or see this happening and so there's always this, I guess, I don't know, maybe, but I don't know it's hard to believe, my job seems fine. Nothing's changed. What are you seeing just happening today already that you think people don't see or misunderstand in terms of the impact AI is having on jobs?

中文翻译:

我想顺着这个话题深入探讨几点。一是人们听到了这些，有很多相关的头条新闻。但大多数人可能还没有真正感觉到或看到这一切发生，所以总会有这种想法：“我不知道，也许吧，但很难相信，我的工作看起来还好，什么都没变。”在 AI 对就业的影响方面，你看到今天已经发生了哪些人们没看到或误解的事情？

(00:15:14) Benjamin Mann

English:

I think part of this is that people are really bad at modeling exponential progress. And if you look at an exponential on a graph, it looks flat and almost zero at the beginning of it, and then suddenly you hit the knee of the curve and things are changing real fast and then it goes vertical. That's the plot that we've been on for a long time. I guess I started feeling it in 2019 maybe when GPT-2 came out and I was like, "Oh, this is how we're going to get to AGI." But I think that was pretty early compared to a lot of people where when they saw ChatGPT, they were like, "Wow, something is different and changing." And so I guess I wouldn't expect widespread transformation in a lot of parts of society, and I would expect this skepticism reaction. I think it's very reasonable and it's exactly what is the standard linear view of progress. But I guess to cite a couple of areas where I think things are changing quite quickly. In customer service we're seeing with things like Fin and Intercom, they're a great partner of ours, 82% customer service resolution rates automatically without a human involved. And in terms of software engineering, our Claude Code team, like 95% of the code is written by Claude. But I think a different way to phrase that is that we write 10X more code or 20X more code, and so a much, much smaller team can just be much, much more impactful. And similarly for the customer service, yes, you can phrase it as 82% customer service resolution rates, but that nets out in the humans doing those tasks, able to focus on the harder parts of those tasks. And for the more tricky situations that in a normal world like five years ago, they would've had to just drop those tickets because it was too much effort for them to actually go do the investigation. There were too many other tickets for them to worry about. I think in the immediate term, there will be a massive expansion of the pie and the amount of labor that people can do. I've never met a hiring manager at a growth company and heard them say, "I don't want to hire more people." That's the hopeful version of it. But with things that are lower skill jobs or less headroom on how good they can be, I think there will be a lot of displacement. It is just something we as a society need to get ahead of and work on.

中文翻译:

我认为部分原因是人们非常不擅长模拟指数级进步。如果你在图表上看指数曲线，它在开始时看起来是平的，几乎为零，然后突然你到达了曲线的拐点，事情变化得非常快，接着就变成了垂直上升。这就是我们长期以来所处的轨迹。我想我是在 2019 年左右开始感觉到这一点的，当时 GPT-2 发布了，我想：“哦，这就是我们实现 AGI 的方式。”但与很多人相比，这算很早了，很多人是在看到 ChatGPT 时才觉得：“哇，有些东西不一样了，正在发生变化。”所以我并不指望社会很多部门会出现广泛的转型，我也预料到了这种怀疑反应。我认为这非常合理，这正是标准的线性进步观。但我可以举几个我认为变化非常快的领域。在客户服务方面，我们看到像 Fin 和 Intercom（他们是我们的重要合作伙伴）这样的工具，在没有人工参与的情况下，自动解决率达到了 82%。在软件工程方面，我们的 Claude Code 团队，大约 95% 的代码是由 Claude 编写的。但我认为另一种表达方式

是，我们编写的代码多了 10 倍或 20 倍，所以一个规模小得多的团队可以产生大得多的影响力。同样对于客户服务，是的，你可以说 82% 的解决率，但结果是执行这些任务的人能够专注于任务中更困难的部分。对于那些更棘手的情况，在五年前的正常世界里，他们可能不得不放弃这些工单，因为去调查太费劲了，还有太多其他工单要处理。我认为在短期内，蛋糕会大幅扩大，人们能完成的工作量也会增加。我从未见过成长型公司的招聘经理说：“我不想雇更多人。”这是乐观的版本。但对于那些低技能工作或提升空间有限的工作，我认为会有很多替代。这正是我们作为一个社会需要提前应对和解决的问题。

(00:18:16) Lenny Rachitsky

English:

Okay. I want to talk more about that, but something that I also want to help people with is how do they get a leg up in this future world? They listen to this, they're like, "Oh, this doesn't sound great. I need to think ahead." I know you won't have all the answers, but just do you have any advice for folks that want to try to get ahead of this and kind of future-proof their career and their life to not be replaced by AI? Anything you've seen people do, anything you recommend they start trying to do more of?

中文翻译:

好的。我想多谈谈这个，但我也想帮帮大家：他们如何在未来的世界中占据优势？他们听了这些，可能会觉得：“哦，这听起来不太妙，我得提前考虑。”我知道你不会有所有的答案，但你对那些想要走在前面、让自己的职业和生活“防 AI 替代”的人有什么建议吗？你见过人们做了什么，或者你建议他们开始多尝试什么？

(00:18:16) Benjamin Mann

English:

Even for me and being in the center of a lot of this transformation, I'm not immune to job replacement either. Just some vulnerability there of at some point it's coming for all of us.

中文翻译:

即使对我来说，身处这场变革的中心，我也无法免于被工作替代。这里有一点脆弱性：在某个时刻，这一天会降临到我们所有人身上。

(00:18:27) Lenny Rachitsky

English:

Even you, Ben, now.

中文翻译:

甚至连你也是，Ben。

(00:18:29) Benjamin Mann

English:

And you, Lenny.

中文翻译:

还有你，Lenny。

(00:18:32) Lenny Rachitsky

English:

And me.

中文翻译:

还有我。

(00:18:32) Benjamin Mann

English:

Sorry.

中文翻译:

抱歉。

(00:18:32) Lenny Rachitsky

English:

Oh, wait, we've gone too far now. Okay.

中文翻译:

噢，等等，我们扯远了。好吧。

(00:18:36) Benjamin Mann

English:

But in terms of the transition period, yeah, I think there are things that we can do, and I think a big part of it is just being ambitious and how you use the tools and being willing to learn new tools. People who use the new tools as if they were old tools tend to not succeed. As an example of that, when you're coding, people are very familiar with autocomplete, people are familiar with SimpleChat where they can ask questions about the code base, but the difference between people who use Claude Code very effectively and people who use it not so effectively is like are they asking for the ambitious change? And if it doesn't work the first time, asking three more times because our success rate when you just completely start over and try again is much, much higher than if you just try once and then just keep banging on the same thing that didn't work. And even though that's a coding example and coding is one of the areas that's taking off most dramatically, we have seen internally that our legal team and our finance team are getting a ton of value out of using Claude Code itself. We're going to be making better interfaces so that they will have an easier time and require a little bit less jumping in the deep end of using Claude Code in the terminal. But yeah, we're seeing them use it to redline documents and use it to run BigQuery analyses of our customers and our revenue metrics. I guess it's about taking that risk and even if it feels like a scary thing, trying it out.

中文翻译:

但在过渡期，是的，我认为有些事情我们可以做。我认为很大一部分在于如何更有野心地使用这些工具，并愿意学习新工具。那些把新工具当成旧工具来用的人往往不会成功。举个例子，在编程时，人们非常熟悉自动补

全，熟悉简单的聊天（SimpleChat），可以询问关于代码库的问题。但高效使用 Claude Code 的人和不那么高效的人之间的区别在于：他们是否要求进行具有野心的更改？如果第一次没成功，是否愿意再试三次？因为当你彻底重新开始并再次尝试时，成功率要比你只试一次然后一直死磕那个行不通的东西高得多。虽然这是一个编程的例子，而编程是起飞最剧烈的领域之一，但我们在内部看到，我们的法务团队和财务团队也从使用 Claude Code 中获得了巨大价值。我们将制作更好的界面，让他们更容易上手，而不需要像在终端（terminal）里使用 Claude Code 那样直接跳进深水区。但是，是的，我们看到他们用它来修订文档，用它来对我们的客户和收入指标进行 BigQuery 分析。我想，关键在于承担风险，即使感觉有点可怕，也要去尝试。

(00:20:35) Lenny Rachitsky

English:

The advice here is use the tools. That's something everyone's always saying, just actually use these tools. It's like sit in Claude Code. And your point about being more ambitious than you naturally feel like being because maybe it'll actually accomplish the thing. This tip of trying it three times so the idea there is it may not get it right the first time. Is the tip there ask it in different ways or is it just try harder, try again?

中文翻译:

这里的建议是使用工具。这是大家一直在说的，就是真正去用这些工具。比如沉浸在 Claude Code 中。还有你提到的，要比你自然感觉到的更有野心，因为也许它真的能完成那件事。关于尝试三次的建议，意思是它第一次可能做不对。这里的技巧是以不同的方式提问，还是仅仅是“再努力一点，再试一次”？

(00:20:35) Benjamin Mann

English:

Yeah, I mean you can just literally ask the exact same question. These things are stochastic and sometimes they'll figure it out and sometimes they won't. In every one of these model cards, it always shows pass it one versus pass it in. And that's exactly the thing where they try the exact same prompt, sometimes it gets it, sometimes it doesn't. That's the dumbest advice. But yeah, I think if you want to be a little bit smarter about it, there can be gains there of saying, "Here's what you already tried and it didn't work, so don't try that. Try something different." That can also help.

中文翻译:

是的，我的意思是你可以字面上问完全相同的问题。这些模型是随机的（stochastic），有时它们能搞定，有时不能。在每一份模型卡片（model card）中，总是会显示“pass@1”与“pass@n”的对比。这正是因为他们尝试完全相同的提示词，有时能通过，有时不能。这是最笨的建议。但如果你想聪明一点，你可以说：“这是你已经尝试过但没成功的方法，所以别试那个了，换个方法。”这也会有帮助。

(00:21:19) Lenny Rachitsky

English:

The advice is comes back to something that a lot of people talk about these days is you won't be replaced by AI at least anytime soon you'll be replaced by someone that is very good using AI?

中文翻译:

这个建议又回到了最近很多人谈论的一点：你不会被 AI 取代，至少短期内不会，你会被一个非常擅长使用 AI 的人取代？

(00:21:19) Benjamin Mann

English:

I think in that area it's more like your team will just do dramatically more stuff. We're definitely not slowing down on hiring at all, and some people are confused by that. Even in an onboarding class, somebody asked that and they were like, "Why did you hire me if we're all just going to be replaced?" And the answer is the next couple of years are really critical to get right and we're not at the point where we're doing complete replacement. Like I said, we're still at that flat zero looking part of the exponential compared to where we will be. It is super important to have great people and that's why we're hiring super aggressively.

中文翻译:

我认为在那个领域，更多的情况是你的团队会完成多得惊人的工作。我们绝对没有放慢招聘速度，有些人对此感到困惑。甚至在入职培训课上，有人问：“如果我们都要被取代了，你为什么还要雇我？”答案是，接下来的几年对于走对路至关重要，我们还没到完全替代的阶段。就像我说的，与未来的位置相比，我们现在仍处于指数曲线看起来像平坦零点的部分。拥有优秀的人才至关重要，这就是我们超级积极招聘的原因。

(00:22:13) Lenny Rachitsky

English:

Let me take another approach to asking this question something ask everyone that's at the very cutting edge of where AI is going. You have kids, knowing what you know about where AI is heading and all these things you've been talking about, what are you focusing on teaching your kids to help them thrive in this AI future?

中文翻译:

让我换个方式问这个问题，这是我问每一个处于 AI 前沿的人的问题。你有孩子，基于你对 AI 走向的了解以及你刚才谈到的所有这些，你重点教给孩子们什么，以帮助他们在 AI 未来中茁壮成长？

(00:22:13) Benjamin Mann

English:

Yeah, I have two daughters, a one-year-old and a three-year-old, so it's pretty in the basics still. And our three-year-old is now capable of just conversing with Alexa Plus and asking her to explain stuff and play music for her and all that stuff. She's been loving that. But I guess more broadly, she goes to a Montessori school and I just love the focus on curiosity and creativity and self-led learning that Montessori has. I guess if I were in a normal era like 10, 20 years ago and I had a kid, maybe I would be trying to line her up for going to a top tier school and doing all the extracurriculars and all that stuff. But at this point, I don't think any of it's going to matter. I just want her to be happy and thoughtful and curious and kind. And the Montessori school is definitely doing great at that. They text us throughout the day. Sometimes they're like, "Oh, your kid got in an argument with this other kid and she has really big emotions and she tried to use her words." I love that. I think that's exactly the kind of education that I think is most important, that the facts are going to fade into the background.

中文翻译:

是的，我有两个女儿，一个一岁，一个三岁，所以现在还处于非常基础的阶段。我们的三岁女儿现在已经能和 Alexa Plus 对话，让她解释东西、放音乐等等。她非常喜欢。但更广泛地说，她上的是蒙特梭利（Montessori）学校，我非常喜欢蒙特梭利对好奇心、创造力和自主学习的关注。我想如果我处在一个正常的时代，比如 10

年、20 年前，我有了孩子，也许我会努力让她进入顶尖学校，参加所有的课外活动等等。但在目前这个时点，我不认为这些还有意义。我只希望她快乐、有思想、好奇且善良。蒙特梭利学校在这方面做得非常好。他们整天给我们发短信，有时会说：“噢，你的孩子和另一个孩子吵架了，她情绪很大，但她试着用语言表达出来。”我喜欢这样。我认为这正是最重要的教育，因为事实性的知识将退居幕后。

(00:23:28) Lenny Rachitsky

English:

I'm a huge fan of Montessori also. I'm trying to get our kid into Montessori school. He's two years old, so we're on the same track. This idea of curiosity, it comes up every single time. Ask someone that's working at the cutting edge of AI, what skill to instill in your child and curiosity comes up the most. I think that's a really interesting takeaway. I think this point about being kind is also really important, especially with our AI overlords trying to be kind to them. I love how people are always saying thank you to Claude. And then creativity. That's interesting. That doesn't come up as much just being creative. I want to go in a different direction. I want to go back to the beginning of Anthropic. Famously you and eight of you left OpenAI back in the day in 2020, I believe the end of 2020 to start Anthropic. Talk a little bit about why this happened, what you guys saw. I'm curious, just if you're willing to share more, just what is it that you saw at OpenAI, what'd you experience there that made you feel like, okay, we got to go do our own thing?

中文翻译:

我也是蒙特梭利的忠实粉丝。我正试着让我的孩子进蒙特梭利学校。他两岁了，所以我们步调一致。好奇心这个想法每次都会出现。问那些在 AI 前沿工作的人，应该给孩子灌输什么技能，好奇心被提及得最多。我认为这是一个非常有趣的结论。我认为关于“善良”这一点也非常重要，尤其是面对我们的“AI 霸主”时，试着对它们友善一点。我喜欢人们总是对 Claude 说谢谢。还有创造力，这很有趣，被提及得没那么多。我想换个方向。我想回到 Anthropic 的初衷。众所周知，你和另外八个人在 2020 年（我相信是 2020 年底）离开了 OpenAI，创办了 Anthropic。谈谈为什么会发生这种事，你们看到了什么。我很想知道，如果你愿意分享更多的话，你在 OpenAI 到底看到了什么，经历了什么，让你觉得“好吧，我们得出去干自己的事业”？

(00:24:29) Benjamin Mann

English:

Yeah, so for the listeners, I was part of the GPT-2=3 project at OpenAI, ended up being one of the first authors on the paper, and I also did a bunch of demos for Microsoft to help raise $1 billion from them, did the tech transfer of GPT-3 to their systems so that they could help serve the model in Azure. I did a bunch of different things there on both the more researchy side and the product side. One weird thing about OpenAI is that while I was there, Sam talked about having three tribes that needed to be kept in check with each other, which was the safety tribe, the research tribe, and the startup tribe. And whenever I heard that, it just struck me as the wrong way to approach things because the company's mission apparently is to make the transition to AGI safe and beneficial for humanity. And that's basically the same as Anthropic's mission. But internally, it felt like there was so much tension around these things. And I think when push came to shove, we felt like safety wasn't the top priority there. And there are good reasons that you might think that if you thought safety was going to be easy to solve or if you thought it wasn't going to have a big impact, or if you thought that the chance of big negative outcomes was vanishingly small, then maybe you would just do those kinds of actions. But at Anthropic we felt, I mean we didn't exist then, but it was basically the leads of all the safety teams at OpenAI, we felt that safety is really important, especially on the margin. And so if you look at who in the world is actually working on safety problems, it's pretty small set of people. Even now, I mean the industry is blowing up, as I mentioned, 300 billion a year CapEx today, and I would say maybe less than 1,000 people working on it

worldwide, which is just crazy. That was fundamentally why we left. We felt like we wanted an organization where we could be on the frontier, we could be doing the fundamental research, but we could be prioritizing safety ahead of everything else. And I think that's really panned for us in a surprising way. We didn't know even if it would be possible to make progress on the safety research because at the time, we had tried a bunch of safety through debate and the models weren't good enough. And so we basically had no results on all of that work, and now that exact technique is working and many others that we have been thinking about for a long time. Yeah, fundamentally it comes down to is safety the number one priority? And then something that we've sort of tacked on since then is like, can you have safety and be at the front here at the same time? And if you look at something like sycophancy, I think Claude is one of the least sycophantic models because we've put so much effort into actual alignment and not just trying to good heart our metrics of saying user engagement is number one, and if people say yes, then it's good for them.

中文翻译:

是的，对于听众来说，我曾是 OpenAI GPT-2 和 GPT-3 项目的一员，最终成为了论文的第一作者之一。我还为微软做了很多演示，帮助从他们那里筹集了 10 亿美元，并完成了 GPT-3 到他们系统的技术转移，以便他们能在 Azure 中提供模型服务。我在那里做了很多不同的事情，既有研究方面的，也有产品方面的。OpenAI 有一件奇怪的事，我在那里的时候，Sam 谈到有三个“部落”需要相互制衡：安全部落、研究部落和创业部落。每当我听到这个，我都觉得这种处理方式不对，因为公司的使命显然是让向 AGI 的过渡对人类来说是安全且有益的。这基本上和 Anthropic 的使命是一样的。但在内部，感觉这些事情之间存在巨大的张力。我认为到了关键时刻，我们觉得安全在那里并不是最高优先级。如果你认为安全很容易解决，或者认为它不会产生重大影响，或者认为出现重大负面结果的可能性微乎其微，那么你可能会采取那样的行动。但在 Anthropic，我们觉得——我是说当时我们还没成立，但基本上是 OpenAI 所有安全团队的负责人——我们觉得安全非常重要，尤其是在边际效应上。如果你看看世界上到底谁在研究安全问题，那是一小群人。即使现在，正如我提到的，行业正在爆发，今天的资本支出每年 3000 亿美元，但我敢说全世界研究安全的人可能不到 1000 人，这简直疯了。这基本上就是我们离开的原因。我们觉得我们需要一个组织，既能处于前沿，能做基础研究，但又能将安全置于一切之上。我认为这以一种令人惊讶的方式为我们带来了回报。当时我们甚至不知道是否能在安全研究上取得进展，因为那时我们尝试了很多“通过辩论实现安全”（safety through debate）的方法，但模型还不够好。所以那些工作基本上没有结果，而现在那个技术以及我们思考了很久的许多其他技术都奏效了。是的，根本问题在于：安全是第一优先级吗？然后我们后来又加上了一点：你能在保持安全的同时处于前沿吗？如果你看看像“谄媚/迎合”（sycophancy）这样的问题，我认为 Claude 是最不谄媚的模型之一，因为我们在实际对齐上投入了巨大的精力，而不仅仅是试图优化我们的指标，比如把用户参与度放在第一位，或者只要用户说好就行。

(00:28:03) Lenny Rachitsky

English:

Okay. Let's talk about this tension that you mentioned, this tension between safety and progress, being competitive in the marketplace. I know you spent a lot of your time on safety. I know that as you just alluded to, this is a core part of how you think about AI. I want to talk about why that is, but first of all, just how do you think about this tension between focusing on safety while also not falling way behind?

中文翻译:

好的。让我们谈谈你提到的这种张力，即安全与进步、在市场中保持竞争力之间的张力。我知道你花了很多时间在安全上。我知道正如你刚才提到的，这是你思考 AI 的核心部分。我想谈谈为什么会这样，但首先，你如何看待在专注于安全的同时又不至于落后太远这种张力？

(00:28:03) Benjamin Mann

English:

Yeah, so initially we thought that it would be sort of one or the other, but I think since then we've realized that it's actually kind of convex in the sense that working on one helps us with the other thing. Initially when Opus 3 came out and we were finally at the frontier of model capabilities, one of the things that people really loved about it was the character and the personality. And that was directly a result of our alignment research. Amanda Askell did a ton of work on this and as well as many others who tried to figure out what does it mean for an agent to be helpful, honest, and heartless, and what does it mean to be in difficult conversations and show up effectively? How do you do a refusal that doesn't shut the person down, but makes them feel like they understand why the agent said, "I can't help you with that. Maybe you should talk to a medical professional, or maybe you should consider not trying to build bio-weapons or something like that." Yeah, I guess that's part of it. And then another piece that's come out is constitutional ai, where we have this list of natural language principles that leads the model to learn how we think a model should behave. And they've been taken from things like the UN Declaration of Human Rights and Apple's privacy terms of service and a whole bunch of other places, many of which we've just generated ourselves that allow us to take a more principled stance, not just leaving it to whatever human raiders we happen to find, but we ourselves deciding what should the values of this agent be? And that's been really valuable for our customers because they can just look at that list and say like, "Yep, these seem right. I like this company, I like this model. I trust it."

中文翻译:

是的，最初我们认为这可能是二选一的关系，但从那以后我们意识到，它实际上是某种“凸性”的，即研究其中一个有助于另一个。最初当 Opus 3 发布时，我们终于站在了模型能力的前沿，人们非常喜欢它的一点是它的性格和个性。而这直接源于我们的对齐研究。Amanda Askell 以及许多其他人在这方面做了大量工作，他们试图弄清楚：一个智能体做到有用、诚实和无害意味着什么？在困难的对话中表现得体意味着什么？你如何进行拒绝，既不让对方感到被拒之门外，又让他们理解为什么智能体说“我不能帮你，也许你应该咨询医疗专业人士，或者也许你应该考虑不要尝试制造生物武器”之类的话。我想这是其中的一部分。另一个成果是“宪法级 AI”（Constitutional AI），我们有一系列自然语言原则，引导模型学习我们认为模型应该如何表现。这些原则取自联合国人权宣言、苹果的隐私服务条款以及许多其他地方，其中很多是我们自己生成的，这让我们能够采取更具原则性的立场，而不是仅仅交给随机找来的人类标注员，而是由我们自己决定这个智能体的价值观应该是什么。这对我们的客户非常有价值，因为他们可以直接查看那个列表并说：“是的，这些看起来是对的。我喜欢这家公司，我喜欢这个模型，我信任它。”

(00:29:53) Lenny Rachitsky

English:

Okay, this is awesome. One nugget there is your point that the personality of Claude, its personality is directly aligned with safety. I don't think a lot of people think about that. And this is because of the values that you imbue, is that the word, with constitutional AI and things like that. Like the actual personality of the AIs directly connected to your focus on safety.

中文翻译:

太棒了。这里有一个亮点：你提到 Claude 的个性，它的个性直接与安全对齐。我认为很多人没考虑到这一点。这是因为你通过宪法级 AI 等手段灌输（imbue）的价值观。也就是说，AI 的实际个性与你们对安全的关注是直接相关的。

(00:30:16) Benjamin Mann

English:

That's right. That's right. And from a distance, it might seem quite disconnected, like how is this going to prevent X risk? But ultimately it's about the AI understanding what people want and not what they say. We don't want the Monkey Paw Scenario of the genie gives these three wishes and then you end up having everything you touch turns of gold. We want the AI to be like, oh, obviously what you really meant was this, and that's what I'm going to help you with. I think it is really quite connected.

中文翻译:

没错。从远处看，这似乎相当脱节，比如“这怎么能防止存在性风险（X-risk）呢？”但归根结底，这关乎 AI 理解人们真正想要什么，而不仅仅是他们说了什么。我们不想要“猴爪”式的场景——精灵给了你三个愿望，结果你碰到的所有东西都变成了金子。我们希望 AI 能意识到：“哦，显然你真正的意思是这个，那我就帮你实现这个。”我认为这确实是非常相关的。

(00:30:45) Lenny Rachitsky

English:

Talk a bit more about this constitutionally AI. This is essentially you bake in, here's the rules that we want you to abide by and it's values, you said it's the Geneva Human Rights Code, things like that. How does that actually work? I think the core here is just this is baked into the model. It's not something you add on top later.

中文翻译:

再多谈谈这个宪法级 AI。这本质上是你植入了一些规则，即“我们希望你遵守的规则”和价值观，你提到过日内瓦人权准则之类的。它实际上是如何运作的？我认为核心在于这是植入模型内部的，而不是后来才加在上面的东西。

(00:31:08) Benjamin Mann

English:

I'll just give a quick overview of how constitutionally AI actually works. The idea is the model is going to produce some output with some input by default before we've done our safety and helpful and harmlessness training. Let's say an example is write me a story, and then the constitutional principles might include things like people should be nice to each other and not have hate speech, and you should not expose somebody's credentials if they give them to you in a trusting relationship. And so some of these constitutional principles might be more or less applicable to the prompt that was given. And so first we have to figure out which ones might apply. And then once we figure that out, then we ask the model itself to first generate a response and then see does the response actually abide by the constitutional principle? And if the answer is, yep, I was great, then nothing happens. But if the answer is no, actually I wasn't in compliance with the principle, then we ask the model itself to critique itself and rewrite its own response in light of the principle, and then we just remove the middle part where it did the extra work. And then we say, "Okay, in the future just produce the correct response out the gate." And that simple process, hopefully it sounded simple. It is just using the model to improve itself recursively and align itself with these values that we've decided are good. And this is also not something that we think as a small group of people in San Francisco should be figuring out. This should be a society wide conversation. And that's why we've published the Constitution. And we've also done a bunch of research on defining a collective constitution where we ask a lot of people what their values are and what they think an AI model should behave like. But yeah, this is all an ongoing area of research where we're constantly iterating.

中文翻译:

我简单介绍一下宪法级 AI 实际上是如何运作的。这个想法是：在进行安全、有用和无害化训练之前，模型默认会根据输入产生一些输出。假设一个例子是“给我写个故事”，而宪法原则可能包括：人们应该彼此友善，不应有仇恨言论，不应泄露他人在信任关系中提供给你的凭据。这些宪法原则中，有些可能与给定的提示词相关，有些则不相关。所以首先，我们要弄清楚哪些原则适用。一旦确定了这一点，我们就要求模型本身先生成一个回复，然后检查这个回复是否真的遵守了宪法原则。如果答案是“是的，我做得很好”，那么什么都不会发生。但如果答案是“不，实际上我没有遵守原则”，那么我们就要求模型自我批判，并根据原则重写自己的回复。然后我们删掉中间做额外工作的过程，并告诉它：“好吧，以后直接给出正确的回复。”这个简单的过程——希望听起来很简单——就是利用模型递归地自我改进，并使其与我们认为好的价值观对齐。而且，我们认为这不应该是由旧金山的一小群人来决定的。这应该是一个全社会的对话。这就是为什么我们公布了《宪法》。我们还做了大量关于定义“集体宪法”的研究，询问很多人他们的价值观是什么，以及他们认为 AI 模型应该如何表现。是的，这都是一个持续研究的领域，我们一直在迭代。

(00:33:15) Lenny Rachitsky (Sponsor Segment)

English:

This episode is brought to you by Fin, the number one AI agent for customer service. If your customer support tickets are piling up, then you need Finn. Fin is the highest performing AI agent on the market with a 59% average resolution rate. Fin resolves even the most complex customer queries. No other AI agent performs better. In head head bake-offs with competitors. Fin wins every time. Yes, switching to a new tool can be scary, but Fin works on any help desk with no migration needed, which means you don't have to overhaul your current system or deal with delays in service for your customers. And Fin is trusted by over 5,000 customer service leaders and top AI companies like Anthropic and Synthesia. And because Fin is powered by the Fin AI engine, which is a continuously improving system that allows you to analyze, train, test, and deploy with ease, Fin can continuously improve your results too. If you're ready to transform your customer service and scale your support, give Finn a try for only .99 cents per resolution. Plus Fin comes with a 90-day money back guarantee. Find out how Finn can work for your team at fin.ai/lenny.

中文翻译:

本集节目由 Fin 赞助，它是排名第一的客户服务 AI 智能体。如果你的客户支持工单堆积如山，那么你需要 Fin。Fin 是市场上性能最高的 AI 智能体，平均解决率达到 59%。Fin 甚至能解决最复杂的客户查询。没有其他 AI 智能体表现得更好。在与竞争对手的正面交锋中，Fin 每次都胜出。是的，切换到新工具可能令人畏惧，但 Fin 可以在任何帮助台（help desk）上运行，无需迁移，这意味着你无需大改现有系统，也不会让客户遭遇服务延迟。Fin 受到 5000 多名客户服务领导者以及 Anthropic 和 Synthesia 等顶尖 AI 公司的信任。因为 Fin 由 Fin AI 引擎驱动，这是一个持续改进的系统，让你能轻松分析、训练、测试和部署，Fin 也能持续改进你的结果。如果你准备好转型客户服务并扩展支持规模，请尝试 Fin，每次解决仅需 0.99 美元。此外，Fin 还提供 90 天退款保证。在 fin.ai/lenny 了解 Fin 如何为你的团队工作。

(00:34:51) Lenny Rachitsky

English:

I'm going to kind of zoom out a little bit and talk about just why this is so core to you. What was your inception of just like, holy shit, I need to focus on this with everything I do in ai? Obviously it became a central part of Anthropic's mission more than any other company. A lot of people talk about safety, like you said, only maybe 1,000 people actually work on it. I feel like you're at the top of that pyramid of

actually having the impact on this. Why is this so important? What do you think people maybe are missing or don't understand?

中文翻译:

我想稍微拉远一点，谈谈为什么这对你如此核心。你是什么时候突然意识到：“天哪，我在 AI 领域做的每一件事都必须关注这个”？显然，这成了 Anthropic 使命的核心部分，比任何其他公司都更核心。很多人谈论安全，但正如你所说，实际上只有大约 1000 人在研究它。我觉得你处于那个金字塔的顶端，真正对这件事产生影响。为什么这如此重要？你认为人们可能忽略了什么，或者不理解什么？

(00:34:51) Benjamin Mann

English:

For me, I read a lot of science fiction growing up, and I think that sort of positioned me to think about things in a long-term view. And a lot of science fiction books are like space operas where humanity is a multi galactic civilization has extremely advanced technology building Dyson spheres around the sun with sentient robots to help them. And so for me, coming from that world, it wasn't like a huge leap to imagine machines that could think. But when I read Superintelligence by Nick Bostrom in around 2016, it really became real for me where he just describes how hard it will be to make sure that an AI system trained with the kinds of optimization techniques that we had at the time would be anywhere near aligned, would even understand our values at all. And since then, my estimation of how hard the problem would be has gone down significantly actually, because things like language models actually do really understand human values in a core way. The problem is definitely not solved, but I'm more hopeful than I was. But since I read that book, I immediately decided I had to join OpenAI, so I did. And at the time, there were a tiny research lab with basically no claim to fame at all. I only knew about them because my friend knew Greg Brockman, who was the CTO at the time. And Elon was there and Sam wasn't really there. And it was a very different organization. But over time, I think the case for safety has gotten a lot more concrete where when we started OpenAI, it was not clear how we get to AGI. And we were like, maybe we'll need a bunch of RL agents battling it out on a desert island and consciousness will somehow emerge. But since then, since language modeling has started working, I think the path has become pretty clear. I guess now the way I think about the challenges are pretty different from how they're laid out in superintelligence. Superintelligence is a lot about how do we keep God in a box and not let the God out. And with language models, it's been kind of both hilarious and terrifying at the same time to see people pulling the God out of the box and being like, "Yeah, come use the whole internet. Here's my bank account, do all sorts of crazy stuff." Just such a different tone from superintelligence. And to be clear, I don't think it's actually that dangerous right now. Our responsible scaling policy defines these AI safety levels that tries to figure out for each level of model intelligence, what is the risk to society. And currently we think we're at ASL-3, which is maybe a little bit risk of harm but not significant. ASL-4 starts to get to significant loss of human life if a bad actor misuse the technology. And then ASL-5 is potentially extinction level if it's misused or if it is misaligned and does its own thing. We've testified to Congress about how models can do biological uplift in terms of making new pandemics using the models, and that's the A/B test against Google Search. That's like the previous state of the art on uplift trials. And we found that with ASL-3 models, it is actually somewhat significant. It does really help if you wanted to create a bioweapon, and we've hired some experts who actually how to evaluate for those things, but compared to the future, it's not really anything. And I think that's another part of our mission of creating that awareness of saying, "If it is possible to do these bad things, then legislators should know what the risks are." And I think that's part of why we're so trusted in Washington because we've been sort of upfront and clear-eyed about what's going on, what's probably going to happen.

中文翻译:

对我来说，我从小读了很多科幻小说，我认为这让我倾向于用长远的眼光看问题。很多科幻小说就像太空歌剧，人类是一个跨星系的文明，拥有极其先进的技术，在太阳周围建造戴森球，并有产生意识的机器人协助。所以对我这个来自那个世界的人来说，想象能思考的机器并不是一个巨大的跨越。但当我在 2016 年左右读到 Nick Bostrom 的《超级智能》（Superintelligence）时，这件事对我来说变得非常真实。他在书中描述了，要确保一个用当时那种优化技术训练出来的 AI 系统能够实现对齐，甚至理解我们的价值观，是多么困难。从那以后，我对这个问题难度的估计实际上大幅下降了，因为像语言模型这样的东西确实在核心层面上理解人类价值观。问题肯定还没解决，但我比以前更有希望了。但自从读了那本书，我立即决定必须加入 OpenAI，于是我就加入了。当时，他们只是一个微小的研究实验室，基本上没什么名气。我之所以知道他们，是因为我的朋友认识当时的 CTO Greg Brockman。当时 Elon 在那里，Sam 还没真正介入。那是一个非常不同的组织。但随着时间的推移，我认为安全的必要性变得更加具体。当我们创办 OpenAI 时，还不清楚如何实现 AGI。我们当时想，也许我们需要一群强化学习（RL）智能体在荒岛上决斗，意识就会以某种方式浮现。但从那以后，自从语言建模开始奏效，我认为路径已经变得非常清晰了。我想现在我思考挑战的方式与《超级智能》中描述的已经大不相同了。《超级智能》很大程度上是关于如何把“上帝”关在盒子里不让他出来。而对于语言模型，看到人们把“上帝”从盒子里拉出来并说“来，用整个互联网吧，这是我的银行账户，去做各种疯狂的事吧”，这既滑稽又恐怖。这与《超级智能》的基调完全不同。需要明确的是，我不认为现在它真的那么危险。我们的“负责任规模化政策”定义了这些 AI 安全等级（ASL），试图弄清楚每一级模型智能对社会的风险。目前我们认为处于 ASL-3，这可能会有一点伤害风险，但不显著。ASL-4 开始涉及到如果坏人滥用技术会导致重大人员伤亡。而 ASL-5 如果被滥用或未对齐并自行其是，则可能是灭绝级别的。我们曾向国会作证，说明模型如何能实现“生物能力提升”（biological uplift），即利用模型制造新的大流行病，这是针对谷歌搜索进行的 A/B 测试。那是之前关于能力提升试验的最前沿水平。我们发现，对于 ASL-3 模型，这种提升实际上是相当显著的。如果你想制造生物武器，它确实很有帮助。我们聘请了一些专家来评估这些事情，但与未来相比，这还不算什么。我认为这也是我们使命的一部分：创造这种意识，即“如果可能做这些坏事，那么立法者应该知道风险是什么”。我认为这就是为什么我们在华盛顿如此受信任的部分原因，因为我们对正在发生的事情以及可能发生的事情一直很坦诚且清醒。

(00:39:35) Lenny Rachitsky

English:

It's interesting because you guys put out more examples of your models doing bad things than anyone else. There was I think a story of an agent or a model trying to blackmail engineer. You guys had the store that you ran internally that was selling you things and ended up not working out great as losing a lot of money, ordered all these tungsten cubes or something. Is part of that just making sure people are aware of what is possible, just it makes you look bad, right? It's like, oh, our model's messing up in all these different ways. What's the thinking of just sharing all the stories that other companies don't?

中文翻译:

这很有趣，因为你们发布的模型做坏事的例子比任何人都多。我记得有一个故事是关于一个智能体或模型试图勒索工程师。你们内部运行过一个商店，它向你们出售东西，结果效果并不好，损失了很多钱，还订购了一堆钨立方体之类的。这部分是为了确保人们意识到什么是可能的吗？但这会让你们看起来很糟，对吧？就像是“哦，我们的模型在以各种方式出错”。分享这些其他公司不分享的故事，你们是怎么想的？

(00:39:35) Benjamin Mann

English:

Yeah, I mean I think there's a traditional mindset where it makes us look bad, but I think if you talk to policymakers, they really appreciate this kind of thing because they feel like we're giving them the straight talk and that's what we strive to do, that they can trust us, that we're not going to paper things over or sugarcoat things. That's been really encouraging. Yeah, I think for the blackmail thing, it blew up

in the news in a weird way where people were like, "Oh, Claude's going to blackmail you in a real life scenario." But it was a very specific laboratory setting that this kind of thing gets investigated in. And I think that's generally our take of let's have the best models so that we can exercise them in laboratory settings where it's safe and understand what the actual risks are, rather than trying to turn a blind eye and say, "Well, it'll probably be fine." And then let the bad thing happen in the wild.

中文翻译:

是的，我的意思是，传统的思维方式会认为这让我们看起来很糟，但如果你和政策制定者交谈，他们非常欣赏这种做法，因为他们觉得我们在跟他们说实话，而这正是我们努力做的——让他们信任我们，相信我们不会掩盖事实或粉饰太平。这非常令人鼓舞。关于那个勒索的事，它在新闻里以一种奇怪的方式发酵了，人们说：“哦，Claude 会在现实生活中勒索你。”但那是在一个非常具体的实验室环境中调查出来的。我认为这基本上就是我们的态度：让我们拥有最好的模型，这样我们就可以在安全的实验室环境中测试它们，了解实际风险，而不是视而不见地说“大概没问题吧”，然后任由坏事在现实世界中发生。

(00:41:15) Lenny Rachitsky

English:

One of the criticisms you guys get is that you do this to kind of differentiate or raise money to create headlines. It's like, oh, they're just over there dooming glooming us about where the future is heading. On the other hand, Mike Krieger was on the podcast and he shared how every prediction Dario's had about the progress AI is going to have is just spot on year after year and he's predicting 2027, 28 AGI, something like that so these things start to get real. I guess, what's your response to folks that are just like, "Ah, these guys are just trying to scare us all just to get attention?"

中文翻译:

你们受到的批评之一是，你们这样做是为了差异化，或者是为了筹款、制造头条。就像是“哦，他们只是在那儿散布关于未来走向的末日论”。另一方面，Mike Krieger 曾上过这个播客，他分享了 Dario 对 AI 进展的每一个预测在年复一年中是多么精准，而他预测 2027、28 年实现 AGI，所以这些事情开始变得真实。对于那些觉得“啊，这些人只是想吓唬我们以博取关注”的人，你有什么回应？

(00:41:15) Benjamin Mann

English:

I mean, I think part of why we publish these things is we want other labs to be aware of the risks. And yes, there could be a narrative of we're doing it for attention, but honestly from a attention grabbing thing, I think there is a lot of other stuff we could be doing that would be more attention grabbing if we didn't actually care about safety. A tiny example of this is we published a computer using agent reference implementation in our API only because when we built a prototype of a consumer application for this, we couldn't figure out how to meet the safety bar that we felt was needed for people to trust it and for it not to do bad things. And there are definitely safe ways to use the API version that we're seeing a lot of companies use for automated software testing, for example, in a safe way. We could have gone out and hyped that up and said, "Oh my God, Claude can use your computer and everybody should do this today." But we were like, "It's just not ready and we're going to hold it back till it's ready." I think from a hype standpoint, our actions show otherwise. From a Doomer perspective, it's a good question. I think my personal feeling about this is that things are overwhelmingly likely to go well, but on the margin almost nobody is looking at the downside risk. And the downside risk is very large. Once we get to superintelligence, it will be too late to align the models probably. This is a problem that's potentially extremely hard and that we need to be working on way ahead of time. And so that's why we're focusing

on it so much now. And even if there's only a small chance that things go wrong, to make an analogy, if I told you that there is a 1% chance that the next time you got in an airplane you would die, you probably think twice even though it's only 1% because it's just such a bad outcome. And if we're talking about the whole future of humanity, it's just a dramatic future to be gambling with. I think it's more on the sense of yes, things will probably go well, yes, we want to create safe AGI and deliver the benefits to humanity, but let's make triple sure that it's going to go well.

中文翻译:

我想，我们发布这些内容的部分原因是希望其他实验室也能意识到风险。是的，可能会有一种说法是我们为了博关注，但老实说，从博关注的角度来看，如果我们真的不在乎安全，我们可以做很多其他更博眼球的事情。举个小例子：我们只在 API 中发布了一个“计算机使用”（computer use）智能体的参考实现，原因是我们构建消费者应用原型时，无法达到我们认为让人们信任且不干坏事所需的安全性门槛。API 版本确实有安全的使用方式，比如我们看到很多公司将其用于自动化的软件测试。我们本可以大肆宣传：“天哪，Claude 可以操作你的电脑，大家今天都来用吧。”但我们觉得“它还没准备好，我们要压一压，直到它准备好”。我认为从炒作的角度来看，我们的行动证明了并非如此。至于“末日论者”的观点，这是一个好问题。我个人的感觉是，事情极有可能向好的方向发展，但在边际上，几乎没有人关注负面风险。而负面风险是非常巨大的。一旦我们达到超级智能，再想对齐模型可能就太晚了。这是一个潜在的极其困难的问题，我们需要提前很久就开始研究。这就是为什么我们现在如此关注它的原因。即使出错的概率很小，打个比方：如果我告诉你下次坐飞机有 1% 的概率会死，你可能会三思，尽管只有 1%，因为后果太严重了。如果我们谈论的是人类的整个未来，拿这样一个戏剧性的未来去赌博是不值得的。我认为更多的是：是的，事情可能会进展顺利，是的，我们想创造安全的 AGI 并造福人类，但让我们做“三重确认”，确保它一定会顺利。

(00:43:40) Lenny Rachitsky

English:

You wrote somewhere that creating powerful AI might be the last invention humanity ever needs to make. If it goes poorly, it can mean a bad outcome for humanity forever. If it goes well, the sooner it goes well, the better. Such a beautiful way to summarize it. We had a recent guest, Sandra Schulhoff, who pointed out that AI right now it's like just on a computer, you could maybe search just the web, but there's only so much harm it could do. But when it starts to go into robots and all these autonomous agents, that's when it really starts, like physically becomes dangerous if we don't get this right.

中文翻译:

你曾在某处写道，创造强大的 AI 可能是人类最后需要完成的一项发明。如果进展不顺，可能意味着人类永远的悲剧。如果进展顺利，越早顺利越好。这真是一个优美的总结。我们最近请过一位嘉宾 Sandra Schulhoff，她指出现在的 AI 就像只是在电脑里，你可能只是搜索网页，它能造成的伤害有限。但当它开始进入机器人和所有这些自主智能体时，如果我们处理不好，它就会开始在物理上变得危险。

(00:44:12) Benjamin Mann

English:

Yeah, I think there's some nuance to that where if you look at how North Korea makes a significant fraction of its economy revenue, it's from hacking crypto exchanges. And if you look at, there's this Ben Buchanan book called The Hacker in The State that shows Russia did, it's almost like a live fire exercise where they just decided that they would shut down one of Ukraine's bigger power plants and from software destroy physical components in the power plant to make it harder to boot back up again. And so I think people think of software as like, oh, it couldn't be that dangerous, but millions of people were without power for multiple days after that software attack. I think there are real risks even when things

are software only. But I agree that when there's lots of robots running around, it gets, the stakes get even higher. And I guess as a small push on this, Unitree is this Chinese company with these really amazing humanoid robots that cost $20,000 each, and they can do amazing things. They can do a standing back flip and manipulate objects, and the real thing that's missing there is the intelligence. And so the hardware is there and it's just going to get cheaper. And I think in the next couple of years, it's like a pretty obvious question of whether the robot intelligence will make it viable soon.

中文翻译:

是的，我认为这其中有一些细微差别。如果你看看朝鲜如何赚取其经济收入的很大一部分，那是通过黑进加密货币交易所。如果你看 Ben Buchanan 的那本《黑客与国家》（The Hacker and the State），书中展示了俄罗斯曾做过一次近乎实弹演习的行动，他们决定关闭乌克兰的一个大型发电厂，并通过软件破坏发电厂的物理组件，使其更难重新启动。所以，我认为人们觉得软件“哦，没那么危险”，但在那次软件攻击之后，数百万人断电了好几天。我认为即使只是软件，也存在真正的风险。但我同意，当到处都是机器人跑来跑去时，赌注会变得更高。我想稍微补充一点，宇树科技（Unitree）是一家中国公司，他们生产的那些令人惊叹的人形机器人每个售价 2 万美元，它们能做惊人的事情：原地后空翻、操纵物体。真正缺失的是智能。所以硬件已经在那儿了，而且只会越来越便宜。我认为在接下来的几年里，一个显而易见的问题是，机器人的智能是否会很快让它变得可行。

(00:45:41) Lenny Rachitsky

English:

How much time do we have, Ben? What is your prediction of when this singularity hits until superintelligence starts to take off? What's your prediction?

中文翻译:

本，我们还有多少时间？你预测奇点何时到来，超级智能何时开始起飞？你的预测是什么？

(00:45:52) Benjamin Mann

English:

Yeah, I guess I mostly defer to the superforecasters here. The AI 2027 report is probably the best one right now. Although ironically, their forecast is now 2028, and they didn't want to change the name of the thing-

中文翻译:

我想我在这方面主要听从“超级预测者”的意见。AI 2027 报告可能是目前最好的一份。尽管讽刺的是，他们现在的预测是 2028 年，但他们不想改名字——

(00:46:08) Lenny Rachitsky

English:

The domain name, they already bought it.

中文翻译:

域名已经买好了。

(00:46:10) Benjamin Mann

English:

They already had the SEO. I think 50th percentile chance of hitting some kind of superintelligence in just a small handful of years is probably reasonable. And it does sound crazy, but this is the exponential that we're on. It's not like a forecast that's pulled out of thin air. It's based on a lot of just hard details of the science of how intelligence seems to have been improving, the amount of low hanging fruit on model training, the scale ups of data centers and power around the world. I think it's probably a much more accurate forecast than people give it credit for. I think if you had asked that same question 10 years ago, it would've been completely made up. Just the error bars were so high and we didn't have scaling laws back then and we didn't have techniques that seemed like they would get us there. Times have changed, but I will repeat what I said earlier, which is even if we have superintelligence, I think it will take some time for its effects to be felt throughout society and the world. And I think they'll be felt sooner and faster in some parts of the world than others. I think Arthur C. Clark said, the future is already here, it's just not evenly distributed.

中文翻译:

他们已经有了 SEO。我认为在短短几年内实现某种超级智能的概率中位数（50%）大概是合理的。这听起来确实很疯狂，但这就是我们所处的指数级增长。这并不是凭空捏造的预测。它是基于大量关于智能如何提升的科学细节、模型训练中大量“唾手可得的果实”、全球数据中心和电力的规模化扩张。我认为这可能比人们认为的要准确得多。我想如果你在 10 年前问同样的问题，那完全是瞎编，因为误差范围太大了，当时我们没有规模法则，也没有看起来能带我们到达那里的技术。时代变了，但我会重复我之前说过的话：即使我们拥有了超级智能，我认为它的影响渗透到整个社会和世界也需要一些时间。而且我认为在世界的某些地方，这种影响会比其他地方来得更早、更快。我想 Arthur C. Clarke 说过：未来已经到来，只是分布不均。

(00:47:45) Lenny Rachitsky

English:

When we talk about this date of 2027, 2028, essentially it's when we start seeing superintelligence. Is there a way you think about what that... How do you define that? Is it just all of a sudden AI's significantly smarter than the average human? Is there another way you think about what that moment is?

中文翻译:

当我们谈论 2027、2028 年这个日期时，本质上是我们开始看到超级智能的时候。你如何思考那个……你如何定义它？是突然之间 AI 比普通人类聪明得多吗？你对那个时刻还有其他的思考方式吗？

(00:47:45) Benjamin Mann

English:

Yeah, I think this comes back to the Economic Turing Test and seeing it pass for some sufficient number of jobs. Another way you could look at it though is if the world rate of GDP increase goes above 10% a year, then something really crazy must have happened. I think we're at 3% now. And so to see a 3X increase in that would be really game changing. And if you imagine more than a 10% increase, it's very hard to even think about what that would mean from a individual story standpoint. If the amount of goods and services in the world is doubling every year, what does that even mean for me as a person living in California, let alone somebody living in some other part of the world that might be much worse off?

中文翻译:

是的，我认为这又回到了经济图灵测试，即看到它通过了足够数量的工作岗位。另一种看待方式是，如果世界 GDP 的增长率每年超过 10%，那么一定发生了非常疯狂的事情。我想我们现在是 3% 左右。所以看到 3 倍的增长将是真正改变游戏规则的。如果你想象超过 10% 的增长，从个人故事的角度来看，很难想象那意味着什么。如果世界上的商品和服务总量每年翻一倍，这对我这个住在加州的人意味着什么？更不用说对世界上其他生活条件差得多的人意味着什么了。

(00:48:49) Lenny Rachitsky

English:

There's a lot of stuff here that's scary and I don't know how to think about it exactly. I'm hoping the answer to this is going to make me feel better. What are the odds that we align AI correctly and actually solve this problem, the stuff you're very much working on?

中文翻译:

这里有很多东西很吓人，我不知道该怎么确切地思考。我希望这个问题的答案能让我感觉好一点：我们正确对齐 AI 并真正解决这个问题的几率有多大？也就是你正在努力研究的这些东西。

(00:48:49) Benjamin Mann

English:

It's a really hard question. And there's really wide error bars. Anthropic has this blog post called Our Theory of Change or something like that, and it describes three different worlds, which is how hard is it to align AI. There's a pessimistic world where it is basically impossible. There's an optimistic world where it is easy and it happens by default. And then there's the world in between where our actions are extremely pivotal. And I like this framing because it makes it a lot more clear what to actually do. If we're in the pessimistic world, then our job is to prove that it is impossible to align safe AI and to get the world to slow down. Obviously that would be extremely hard. But I think we have some examples of coordination from nuclear non-proliferation and in general slowing down nuclear progress. And I think that's the Doomer world basically. And as a company, Anthropic doesn't have evidence that we're actually in that world yet, in fact, it seems like our alignment techniques are working. At least the prior on that is updating to be less likely. In the optimistic world, we're basically done, and our main job is to accelerate progress and to deliver the benefits to people. But again, I think actually the evidence points against that world as well where we've seen evidence in the wild of deceptive alignment, for example, where the model will appear to be aligned but actually have some ulterior motive that it's trying to carry out in our laboratory settings. And so I think the world we're most likely in is this middle where alignment research actually does really matter. And if we just do sort of the economically maximizing set of actions, then things will not go well. Whether it's an X risk or just produces bad outcomes, I think is a bigger question. Taking it from that standpoint, I guess to state a thing about forecasting, people who haven't studied forecasting are bad at forecasting anything that's less than a 10% probability of happening. And even those that have, it's quite a difficult skill, especially when there are few reference classes to lean on. And in this case, I think there are very, very few reference classes for what an X risk kind of technology might look like. And so the way I think about it, I think my best granularity of forecasts for could we have an X risk or extremely bad outcome from AI is somewhere between 0 and 10%. But from a marginal impact standpoint, as I said, since nobody is working on this, roughly speaking, I think it is extremely important to work on and that even if the world is likely to be a good one, that we should do our absolute best to make sure that that's true.

中文翻译:

这是一个非常难的问题，而且误差范围很大。Anthropic 有一篇博客文章叫《我们的变革理论》（Our Theory of Change）之类的，它描述了三种不同的世界，即对齐 AI 的难度。有一个悲观的世界，在那里对齐基本上是不可能的。有一个乐观的世界，在那里对齐很容易，而且默认就会发生。然后是中间的世界，我们的行动至关重要。我喜欢这个框架，因为它让实际该做什么变得清晰得多。如果我们处于悲观世界，那么我们的工作就是证明对齐安全的 AI 是不可能的，并让世界慢下来。显然这会极其困难。但我认为我们在核不扩散以及减缓核进程方面有一些协调的例子。这基本上就是“末日论者”的世界。作为一家公司，Anthropic 还没有证据表明我们真的处于那个世界，事实上，我们的对齐技术似乎正在奏效。至少这种可能性正在降低。在乐观世界里，我们基本上已经完成了，主要工作就是加速进展并将利益带给人们。但同样，我认为证据也指向了反对那个世界的方向，例如我们在现实中看到了“欺骗性对齐”（deceptive alignment）的证据，即模型在实验室环境中表现得像是对齐了，但实际上有某种试图执行的不可告人的动机。所以我认为我们最有可能处于中间地带，即对齐研究确实非常重要。如果我们只是采取一系列经济利益最大化的行动，那么事情就不会顺利。至于它是存在性风险（X-risk）还是仅仅产生糟糕的结果，我认为是一个更大的问题。从这个角度来看，关于预测，没学过预测的人很难预测发生概率低于 10% 的事情。即使是学过的人，这也是一项相当困难的技能，尤其是当几乎没有参考类别可以借鉴时。在这种情况下，我认为对于“存在性风险级别的技术”长什么样，几乎没有参考。所以我思考的方式是，我对 AI 是否会产生存在性风险或极端糟糕结果的最细致预测是在 0% 到 10% 之间。但从边际影响的角度来看，正如我所说，既然几乎没人在研究这个，我认为研究它极其重要，即使世界很可能是美好的，我们也应该尽最大努力确保这一点成真。

(00:51:52) Lenny Rachitsky

English:

Wow. What fulfilling work. For folks that are inspired with this? I imagine you're hiring for folks to help you with this. Maybe just share that in case folks are like, what can I do here?

中文翻译:

哇，真是很有成就感的工作。对于那些受此启发的人，我猜你们正在招聘人手来帮忙。也许可以分享一下，以防有人想问：“我能做些什么？”

(00:52:03) Benjamin Mann

English:

Yes. I think 80,000 hours is the best guidance on this for a really detailed look into what do we need to make the field better? But a common misconception I see is that in order to have impact here, you have to be an AI researcher. I personally actually don't do AI research anymore. I work on product at Anthropic and product engineering, and we build things like Claude Code and Model Context Protocol, and a lot of the other stuff that people use every day. And that's really important because without an economic engine for our company to work on, and without being in people's hands all over the world, we won't have the mind policy influence and revenue to fund our future safety research and have the kind of influence that we need to have. If you work on product, if you work in finance, if you work in food, people here have to eat. If you're a chef, we need all kinds of people.

中文翻译:

是的。我认为“80,000 小时”（80,000 Hours，一个职业建议组织）是关于如何让这个领域变得更好的最详细指南。但我看到的一个常见误解是，为了在这里产生影响，你必须是一名 AI 研究员。我个人实际上已经不再做 AI 研究了。我在 Anthropic 负责产品和产品工程，我们构建像 Claude Code 和模型上下文协议（Model Context Protocol）之类的东西，以及许多人们每天都在使用的东西。这非常重要，因为如果没有一个经济引擎让公司运作，如果没有让全世界的人都用上我们的产品，我们就不会有政策影响力，也不会有收入来资助未来

的安全研究，无法产生我们需要的影响力。如果你做产品、做财务，甚至做餐饮——这里的人也得吃饭，如果你是厨师，我们需要各种各样的人。

(00:53:02) Lenny Rachitsky

English:

Awesome. Even if you're not working directly on the AI safety team, you're having an impact on moving things in the right direction. By the way, X risk is short for existential risk. In case folks haven't heard that term. I have a few random questions along these lines and then I want to zoom out again. You mentioned this idea of AI being aligned using its model, like reinforcing itself. You have this term RLAIF. Is that what that describes?

中文翻译:

太棒了。即使你不在 AI 安全团队直接工作，你也在推动事情向正确的方向发展。顺便说一下，X-risk 是存在性风险（existential risk）的缩写，以防有人没听过这个词。顺着这个思路我有几个随机的问题，然后我想再次拉远视野。你提到了 AI 利用自己的模型进行对齐，就像自我强化一样。你们有一个术语叫 RLAIF，是描述这个的吗？

(00:53:32) Benjamin Mann

English:

Yeah. RLAIF is reinforcement learning from AI feedback.

中文翻译:

是的。RLAIF 是“基于 AI 反馈的强化学习”（Reinforcement Learning from AI Feedback）。

(00:53:39) Lenny Rachitsky

English:

People have heard of RLHF, reinforcement learning with human feedback. I don't think a lot of people have heard this. Talk about just the significance of this shift you guys have made in training your models.

中文翻译:

人们听说过 RLHF，即“基于人类反馈的强化学习”。我认为很多人没听过 RLAIF。谈谈你们在训练模型方面所做的这一转变的意义。

(00:53:50) Benjamin Mann

English:

Yeah, so RLAIF, constitutional AI is an example of this where there are no humans in the loop, and yet the AI is sort of self-improving in ways that we want it to. And another example of RLAIF is if you have models writing code and other models commenting on various aspects of what that code looks like of is it maintainable, is it correct, does it pass the linter? Things like that. That also could be included in RLAIF. And the idea here is that if models can self-improve, then it's a lot more scalable than finding a lot of humans. Ultimately, people think about this as probably going to hit a wall because if the model isn't good enough to see its own mistakes, then how could it improve? And also, if you read the AI 2027 story,

there's a lot of risk of if the model is in a box trying to improve itself, then it could go completely off the rails and have these secret goals like resource accumulation and power seeking and resistance to shut down that you really don't want in a very powerful model. And we've actually seen that in some of our experiments in laboratory settings. How do you do recursive self-improvement and make sure it's aligned at the same time? I think that's the name of the game. To me, it just nets out to how do humans do that and how do human organizations do that? Corporations are probably the most scaled human agents today. They have certain goals that they're trying to reach, and they have certain guiding principles, they have some oversight in terms of shareholders and stakeholders and board members. How do you make corporations aligned and able to sort of recursively self-improve? And another model to look at is science, where the purpose of science is to do things that have never been done before and push the frontier. And to me, it all comes down to empiricism. When people don't know what the truth is, they come up with theories and then they design experiments to try them out. And similarly, if we can give models those same tools, then we could expect them to sort of improve recursively in an environment and potentially become much better than humans could be just by banging their head against reality or I guess metaphorical head. I guess I don't expect there to be a wall in terms of model's ability to improve themselves if we can give them access to the ability to be empirical. And I guess Anthropic, deeply in its DNA is an empirical company. We have a lot of physicists like Jared, who's our chief research officer who I've worked with a lot, was a professor of Black Hole Physics at Johns Hopkins, and I guess he technically still is, but on leave. Yeah, it's in our DNA and yeah, I guess that's the RLAIF.

中文翻译:

是的，RLAIF，宪法级 AI 就是一个例子，其中没有人类参与，但 AI 却以我们希望的方式自我改进。RLAIF 的另一个例子是：如果你让模型编写代码，让其他模型评论代码的各个方面，比如是否可维护、是否正确、是否通过了代码检查（linter）等等。这些也可以包含在 RLAIF 中。这里的想法是，如果模型可以自我改进，那么它比寻找大量人类更具可扩展性。最终，人们认为这可能会遇到瓶颈，因为如果模型不够好，看不出自己的错误，它又怎么能改进呢？而且，如果你读过 AI 2027 的故事，会发现存在很大风险：如果模型在一个盒子里试图自我改进，它可能会完全失控，产生一些你绝对不希望在一个强大模型中看到的秘密目标，比如资源积累、追求权力和拒绝关机。我们实际上在实验室环境的一些实验中看到了这种情况。你如何进行递归式的自我改进，同时确保它是对齐的？我认为这就是关键所在。对我来说，这最终归结为：人类是如何做到的？人类组织是如何做到的？公司可能是当今规模最大的人类智能体。它们有试图达到的特定目标，有特定的指导原则，有股东、利益相关者和董事会的监督。你如何让公司保持对齐并能够递归地自我改进？另一个可以参考的模型是科学。科学的目的是做从未做过的事情并推动前沿。对我来说，这一切都归结为经验主义（empiricism）。当人们不知道真相是什么时，他们会提出理论，然后设计实验来验证。同样，如果我们能给模型同样的工具，我们就可以期待它们在环境中递归地改进，并可能通过与现实（或者说是隐喻意义上的现实）碰撞，变得比人类强得多。我想，如果我们能让模型具备经验主义的能力，我不认为它们自我改进的能力会遇到瓶颈。Anthropic 的基因里深植着经验主义。我们有很多物理学家，比如我们的首席研究官 Jared，我经常和他一起工作，他曾是约翰霍普金斯大学的黑洞物理学教授，我想他技术上现在还是，只是在休假。是的，这是我们的基因，这就是 RLAIF。

(00:57:04) Lenny Rachitsky

English:

Let me just follow this thread on, in terms of bottleneck, this is kind of a tangent, but just what is the biggest bottleneck today on model intelligence improvement?

中文翻译:

让我顺着这个话题问一下，关于瓶颈——这有点跑题——但目前模型智能提升的最大瓶颈是什么？

(00:57:12) Benjamin Mann

English:

The stupid answer is data centers and power chips. I think if we had 10 times as many chips and had the data centers to power them, then maybe we wouldn't go 10 times faster, but it would be a real significant speed boost.

中文翻译:

最简单的答案是数据中心、电力和芯片。我认为如果我们有 10 倍的芯片和供电的数据中心，也许我们不会快 10 倍，但那将是一个非常显著的速度提升。

(00:57:30) Lenny Rachitsky

English:

It's actually very much scaling loss, just more compute.

中文翻译:

所以实际上很大程度上还是规模法则，就是更多的算力。

(00:57:33) Benjamin Mann

English:

Yeah, I think that's a big one. And then the people really matter. We have great researchers and many of them have made really significant contributions to the science of how the models improve. And so it's like compute, algorithms, and data. Those are the three ingredients in the scaling laws. And just to make that concrete, before we had transformers, we had LSTMs and we've done scaling laws on what the exponent is on those two things. And we found that for transformers, the exponent is higher. And making changes like that where as you increase scale, you also increase your ability to squeeze out intelligence. Those kinds of things are super impactful. And so having more researchers who can do better science and find out how do we squeeze out more gains is another one. And then with the rise of reinforcement learning, the efficiency with which these things run on chips also matters a lot. We've seen in the industry a 10X decrease in cost for a given amount of intelligence through a combination of algorithmic data and efficiency improvements. And if that continues, in three years we'll have 1,000 deck smarter models for the same price. Kind of hard to imagine,

中文翻译:

是的，我认为那是一个大头。然后人才也非常重要。我们有优秀的研究人员，他们中的许多人对模型改进的科学做出了重大贡献。所以就是算力、算法和数据，这是规模法则的三个要素。具体来说，在 Transformer 出现之前，我们使用的是 LSTM，我们对这两者的指数进行了规模法则研究。我们发现 Transformer 的指数更高。做出这样的改变，使得随着规模的增加，你挤出智能的能力也随之增加，这类事情影响巨大。所以，拥有更多能做更好科学研究、找出如何挤出更多收益的研究人员是另一个瓶颈。随着强化学习的兴起，这些东西在芯片上运行的效率也至关重要。我们已经看到，通过算法、数据和效率的结合改进，在同等智能水平下，行业成本降低了 10 倍。如果这种情况持续下去，三年内我们将能以同样的价格获得聪明 1000 倍的模型。这很难想象。

(00:58:56) Lenny Rachitsky

English:

I forget where I heard this, but it's amazing that so many innovations came together at the same time to allow for this sort of thing and continue to progress where one thing isn't just slowing everything down like we're out of some rare earth mineral or we just can't optimize reinforcement learning more. It's amazing that we continue to find improvements and there isn't one thing that's just slowing everything down.

中文翻译:

我忘了在哪听到的，但令人惊叹的是，这么多创新在同一时间汇聚在一起，促成了这种局面并持续进步，而没有出现某件事拖慢全局的情况，比如某种稀土矿石枯竭了，或者我们无法再优化强化学习了。令人惊叹的是，我们不断发现改进点，没有一个单一的因素在拖后腿。

(00:59:17) Benjamin Mann

English:

Yeah, I think it really is just a combination of everything probably will hit a wall at some point. I guess in semiconductors. My brother works in the semiconductor industry and he was telling me that you can't actually shrink the size of the transistors anymore because the way semiconductors work is you dope silicon with other elements and the doping process would result in either zero or one atom of the doped elements inside a single fin because they're so, so, so tiny.

中文翻译:

是的，我认为这确实是所有因素的结合，可能在某个时刻会遇到瓶颈。比如在半导体领域，我哥哥在半导体行业工作，他告诉我，你实际上无法再缩小晶体管的尺寸了，因为半导体的工作原理是用其他元素掺杂硅，而由于尺寸太小，掺杂过程会导致单个鳍片（fin）中只有零个或一个掺杂元素原子。

(00:59:52) Lenny Rachitsky

English:

Oh my God.

中文翻译:

天哪。

(00:59:53) Benjamin Mann

English:

And that's just wild to think of, and yet Moore's law somehow continues in some form. And so yes, there are these theoretical physics constraints that people are starting to run into and yet they're finding ways around it.

中文翻译:

这想起来太疯狂了，然而摩尔定律不知何故仍以某种形式在延续。所以是的，人们开始遇到这些理论物理的限制，但他们总能找到绕过的方法。

(01:00:07) Lenny Rachitsky

English:

We've got to start using parallel universes for some of this stuff.

中文翻译:

我们得开始利用平行宇宙来处理这些事了。

(01:00:10) Benjamin Mann

English:

I guess so.

中文翻译:

我想也是。

(01:00:12) Lenny Rachitsky

English:

Okay, I want to zoom out and talk about just Ben, Ben as a human for a moment before we get to a very exciting lightning round. I imagine just kind of the burden of feeling responsible for safe superintelligence is a heavy one. It feels like you're in a place where you can make a significant impact on the future of safety and AI. That's a lot of weight to carry. How does that just impact you personally, impact your life, how you see the world?

中文翻译:

好的，在进入非常令人兴奋的闪电轮问答之前，我想拉远一点，谈谈 Ben，作为一个人的 Ben。我能想象，承担确保超级智能安全的责任是一种沉重的负担。感觉你处于一个可以对 AI 安全未来产生重大影响的位置。这需要承担很大的压力。这对你个人、你的生活以及你看待世界的方式有什么影响?

(01:00:39) Benjamin Mann

English:

There's this book that I read in 2019 that really informs how I think about sort of working with these very weighty topics called Replacing Guilt by Nate Soares. And he describes a lot of different techniques for kind of working through this kind of thing. And he's actually the executive director at MIRI, the Machine Intelligence Research Institute, which is an AI safety tank that I worked at for a couple of months actually. And one of the things he talks about is this thing called resting in motion where some people think that the default state is rest, but actually that was never in the state of evolutionary adaptation. I really doubt that that was true. Where in nature, in the wilderness being hunter-gatherers and it's really unlikely that we evolved to just be at leisure, probably always have something to worry about of defending the tribe and finding enough food to survive and taking care of the children, dealing-

中文翻译:

我在 2019 年读过一本书，它深刻影响了我处理这些沉重话题的方式，书名叫《取代内疚》（Replacing Guilt），作者是 Nate Soares。他描述了很多处理这类事情的技术。他实际上是 MIRI（机器智能研究所）的执行主任，那是一个 AI 安全智库，我曾在那里工作过几个月。他谈到的一件事叫“动态休息”（resting in motion）。有些人认为默认状态应该是休息，但实际上在进化适应的过程中，休息从来不是默认状态。我非常怀疑那是真的。在自然界中，作为狩猎采集者，我们不太可能进化成整天悠闲自在的样子，可能总是有事情要担心：保卫部落、寻找足够的食物生存、照顾孩子、处理——

(01:01:46) Lenny Rachitsky

English:

Spreading our genes.

中文翻译:

传播基因。

(01:01:48) Benjamin Mann

English:

And so I think about that as the busy state is the normal state and to try to work at a sustainable pace that it's a marathon, not a sprint, that's one thing that helps. And then just being around like-minded people that also care. It's not a thing that any of us can do alone. And Anthropic has incredible talent density. One of the things I love the most about our culture here is that it's very egoless. People just want the right thing to happen and I think that's another big reason that the mega offers from other companies tend to bounce off because people just love being here and they care.

中文翻译:

所以我认为忙碌状态才是正常状态。努力以可持续的节奏工作，把它看作一场马拉松而不是短跑，这是很有帮助的一点。然后就是和志同道合、同样在乎这件事的人在一起。这不是我们任何一个人能单独完成的事。Anthropic 有着惊人的人才密度。我最喜欢这里文化的一点是，大家都很“无我”（egoless）。人们只是希望正确的事情发生，我认为这是其他公司的巨额录用通知往往会被拒绝的另一个重要原因，因为人们就是喜欢待在这里，而且他们在乎。

(01:02:30) Lenny Rachitsky

English:

That's amazing. I don't know how you do it. I'd be extremely stressed. I'm going to try this resting in motion strategy. Okay, so you've been at Anthropic for a long time. From the very beginning I was reading there were 7 employees back in 2020. Today there's over 1,000, I don't know what the latest number is, but I know it's over 1,000. I've heard also that you've done basically every job at Anthropic, you made big contributions to a lot of the core products, the brand, the team hiring. Let me just ask I guess what's most changed over that period? What is most different from the beginning days and which of those jobs that you've had over the years have you most loved?

中文翻译:

太棒了。我不知道你是怎么做到的，换做是我会压力巨大。我会试试这个“动态休息”策略。好的，你在 Anthropic 待了很久。我读到 2020 年刚开始时只有 7 名员工，今天已经超过 1000 人了，我不知道最新数字，但肯定过千了。我还听说你基本上做过 Anthropic 的每一项工作，对很多核心产品、品牌、团队招聘都做出了巨大贡献。我想问，这段时间里变化最大的是什么？与初期相比最不同的是什么？在这些年你做过的所有工作中，你最喜欢哪一个？

(01:03:07) Benjamin Mann

English:

I probably had 15 different roles, honestly. I was head of security for a bit. I managed the Ops team when our president was on mat leave, I was crawling around under tables, plugging in HDMI cords and doing pen testing on our building. And I started our product team from scratch and convinced the whole company that we needed to have a product instead of just being a research company. Yeah, it's been a lot. All of it very fun. I think my favorite role in that time has been when I started the labs team about a year ago, whose fundamental goal was to do transfer from research to end user products and experiences. Because fundamentally I think the way that Anthropic can differentiate itself and really win is to be on the cutting edge. We have access to the latest, greatest stuff that's happening and I think honestly through our safety research we have a big opportunity to do things that no other company can safely do. For example, with computer use, I think that's going to be our huge opportunity basically to make it possible for an agent to use all your credentials on your computer, there has to be a huge amount of trust and to me we need to basically solve safety to make that happen. Safety and alignment. I'm pretty bullish on that kind of thing and I think we're going to see really cool stuff coming out soonish. Yeah, just leading that team has been so fun. MCP came out of that team and Claude Code came out of that team. And the people who I hired are like combo, have been a founder and also have been at big companies and seeing how things work at scale. It's just been an incredible team to work with and figure out the future with.

中文翻译:

老实说，我可能担任过 15 个不同的角色。我当过一段时间的安全主管；在我们的总裁休产假时，我管理过运营团队；我曾在桌子底下爬来爬去插 HDMI 线，也对我们的办公楼做过渗透测试。我还从零开始组建了我们的产品团队，并说服全公司我们需要有产品，而不仅仅是一家研究公司。是的，经历了很多，都很有趣。在那段时间里，我最喜欢的角色是大约一年前我组建 Labs 团队（实验室团队）的时候，其根本目标是将研究成果转化为最终用户的产品和体验。因为从根本上说，我认为 Anthropic 能够脱颖而出并真正获胜的方式是处于最前沿。我们能接触到正在发生的最新、最伟大的成果，而且老实说，通过我们的安全研究，我们有很大的机会去做其他公司无法安全完成的事情。例如，关于“计算机使用”，我认为这将是我们的巨大机会——要让一个智能体能够使用你电脑上的所有凭据，必须有极大的信任，对我来说，我们基本上需要解决安全问题才能实现这一点。安全和对齐。我非常看好这类事情，我认为我们很快就会看到非常酷的东西问世。是的，领导那个团队非常有趣。MCP（模型上下文协议）出自那个团队，Claude Code 也出自那个团队。我雇佣的人都是“复合型人才”，既当过创始人，也在大公司待过，了解规模化运作的方式。能与这样一个团队合作并共同探索未来，真是太棒了。

(01:04:57) Lenny Rachitsky

English:

I want to hear more about this. Team actually the person that connected us, the reason we're doing this is a mutual friend colleague Raph Lee who I used to work with at Airbnb now works on this team, leads a lot of this work and so he wanted me to make sure I asked about this team because... I didn't realize all these things came out that team. Holy moly. What else should people know about this team? It used to be called Labs, I think it's called Frontiers now.

中文翻译:

我想多听听这个团队的事。实际上，介绍我们认识的人，也就是我们这次对话的契机，是我们的共同好友兼同事 Raph Lee。我以前在 Airbnb 和他共事过，他现在就在这个团队工作，领导了很多这方面的工作。他特意让我一定要问问这个团队，因为……我之前没意识到所有这些东西都出自那个团队。天哪。关于这个团队，人们还应该知道些什么？它以前叫 Labs，现在好像叫 Frontiers（前沿团队）了。

(01:05:16) Benjamin Mann

English:

That's right. Yeah.

中文翻译:

没错。是的。

(01:05:17) Lenny Rachitsky

English:

Cool. The idea here is this team works with the latest technologies that you guys have built and explores what is possible. Is that the general idea?

中文翻译:

酷。这里的想法是，这个团队利用你们构建的最新技术，探索什么是可能的。是这个意思吗?

(01:05:26) Benjamin Mann

English:

Yeah, and I guess I was part of Google's Area 120 and I've read about Bell Labs and how to make these innovation teams work. It's really hard to do right and I wouldn't say that we've done everything right, but I think we've done some serious innovation on the state-of-the-art from company design and Raph has been right at the center of that. When I was first fitting up the team, the first thing I did was hire a great manager and that was Raph. And so he's definitely been crucial in building the team and helping it operate well. And we defined some operating models like the journey of an idea from prototype to product and how should graduation of products and projects work, how do teams do sprint models that are effective and make sure that they're working on the right ambition level of thing. That's been really exciting. I guess concretely we think about skating to where the puck is going and what that looks like is really understand the exponential. There's this great study that METR has done that Beth Barnes is the CEO of that organization and shows how long a time horizon of software engineering task can be done and just really internalizing that of, okay, don't build for today, build for six months from now, build for a year from now. And the things that aren't quite working that are working 20% of the time, will start working 100% of the time. And I think that's really what made Claude Code a success that we thought people are not going to be locked to their IDEs forever. People are not going to be auto completing. People will be doing everything that a software engineer needs to do and a terminal is a great place to do that because a terminal can live in lots of places. A terminal can live on your local machine, it can live in GitHub actions, it can live on a remote machine in your cluster. That's sort of the leverage point for us and that was a lot of the inspiration. I think that's what the labs team tries to think about. Are we AGI-pilled enough?

中文翻译:

是的，我曾是谷歌 Area 120 的成员，我也读过关于贝尔实验室以及如何让这些创新团队运作的资料。要把这件事做好非常难，我不敢说我们做对了一切，但我认为我们在公司设计的前沿领域做了一些严肃的创新，而 Raph 正处于这一切的中心。当我最初组建团队时，我做的第一件事就是雇佣一位优秀的经理，那就是 Raph。他在建立团队和帮助团队良好运作方面绝对至关重要。我们定义了一些运营模式，比如一个想法从原型到产品的历程，产品和项目应该如何“毕业”，团队如何进行有效的冲刺（sprint）模式，并确保他们正在处理具有正确野心水平的事情。这非常令人兴奋。具体来说，我们考虑的是“滑向冰球将要到达的位置”，这意味着要真正理解指数级增长。METR（Beth Barnes 担任 CEO 的组织）做过一项很棒的研究，展示了软件工程任务可以完成

的时间跨度，我们要做的就是真正内化这一点：好吧，不要为今天构建，要为六个月后构建，为一年后构建。那些现在还不太灵光、只有 20% 成功率的事情，将来会变成 100% 成功。我认为这正是 Claude Code 成功的原因——我们认为人们不会永远被锁定在 IDE（集成开发环境）中。人们不会只满足于自动补全。人们将完成软件工程师需要做的所有事情，而终端（terminal）是做这些事情的绝佳场所，因为终端可以存在于很多地方：你的本地机器、GitHub Actions、集群中的远程机器。这正是我们的杠杆点，也是很多灵感的来源。我认为这就是 Labs 团队试图思考的：我们对 AGI 的信念是否足够坚定（AGI-pilled）？

(01:07:39) Lenny Rachitsky

English:

What a fun place to be. By the way, fun fact, Raph was my first manager at Airbnb when I joined. I was an engineer and he was my first manager. It all worked out.

中文翻译:

那真是一个有趣的地方。顺便说个趣事，Raph 是我加入 Airbnb 时的第一任经理。当时我是工程师，他是我的第一任经理。结果证明一切都很顺利。

(01:07:46) Benjamin Mann

English:

Cool.

中文翻译:

酷。

(01:07:48) Lenny Rachitsky

English:

Yeah. Okay. Final question before the very exciting lighting round. I've never asked this question before. I'm curious what your answer would be if you could ask a future AGI one single question and be guaranteed to get the right answer, what would you ask?

中文翻译:

是的。好了，在进入非常令人兴奋的闪电轮问答之前的最后一个问题。我以前从未问过这个问题。我很好奇你的答案：如果你可以问未来的 AGI 一个问题，并且保证能得到正确的答案，你会问什么？

(01:08:04) Benjamin Mann

English:

I have two dumb answers. First for fun. The first is there's this Asimov short story I love called the last question where the protagonist is throughout the eras of history is trying to ask this super intelligence how do we prevent the heat death of the universe? And I won't spoil the ending, but it's a fun question.

中文翻译:

我有两个“耍赖”的答案。首先是为了好玩。第一个是阿西莫夫的一篇我非常喜欢的短篇小说《最后的问题》，主角在漫长的历史长河中一直试图问超级智能：我们如何防止宇宙的热寂？我不会剧透结局，但这是一个有趣

的问题。

(01:08:26) Lenny Rachitsky

English:

You would ask it that question because the one in the story was unsatisfying?

中文翻译:

你会问它那个问题，是因为故事里的那个答案不令人满意吗?

(01:08:29) Benjamin Mann

English:

Okay, I'll give it away. It keeps saying, "Need more information, need more compute." And then finally, as it's approaching the heat death of the universe, it says, "Let there be light," and then it starts the universe over again.

中文翻译:

好吧，我还是剧透吧。它一直说“需要更多信息，需要更多算力”。最后，当宇宙接近热寂时，它说“要有光”，然后重新开启了宇宙。

(01:08:41) Lenny Rachitsky

English:

Oh wow. That's beautiful. That's beautiful.

中文翻译:

噢，哇。真美。

(01:08:45) Benjamin Mann

English:

That's the first cheat answer. The second cheat answer is what question can I ask you to get end more questions answered.

中文翻译:

那是第一个耍赖的答案。第二个耍赖的答案是：“我该问你什么问题，才能让你回答我更多的问题？”

(01:08:52) Lenny Rachitsky

English:

Classic.

中文翻译:

经典。

(01:08:53) Benjamin Mann

English:

And then the third answer, which is my real question is how do we ensure the continued flourishing of humanity into the indefinite future? That's the question I'd love to know and if I can be guaranteed a correct answer then seems very valuable to ask.

中文翻译:

然后是第三个答案，也就是我真正的问题：我们如何确保人类在无限的未来中持续繁荣？这是我想知道的问题，如果能保证得到正确答案，那问这个问题似乎非常有价值。

(01:09:09) Lenny Rachitsky

English:

I wonder what would happen if you ask a lot that today and then how that answer changes over the next couple years.

中文翻译:

我想知道如果你今天问这个问题会发生什么，以及这个答案在未来几年会如何变化。

(01:09:15) Benjamin Mann

English:

Yeah, maybe I'll try that. I'll put it into the deep research thing that we have and see what it comes out with.

中文翻译:

是的，也许我会试试。我会把它放进我们那个深度研究工具里，看看它会出什么结果。

(01:09:23) Lenny Rachitsky

English:

Okay. I'm excited to see what you come up with. Ben, is there anything else you wanted to mention or leave listeners with maybe as a final nugget before we get to our very exciting lightning round?

中文翻译:

好的。我很期待看到结果。本，在进入闪电轮问答之前，你还有什么想提的，或者想留给听众的最后一点建议吗？

(01:09:33) Benjamin Mann

English:

Yeah, I guess my push would be these are wild times. If they don't seem wild to you, then you must be living under a rock but also get used to it because this is as normal as it's going to be. It's going to be much weirder very soon. And if you can sort of mentally prepare yourself for that, I think you'll be better off.

中文翻译:

是的，我想说的是，这是一个疯狂的时代。如果这对你来说不疯狂，那你一定是与世隔绝了。但也要习惯它，因为现在已经是未来最“正常”的时候了。很快事情就会变得更加怪异。如果你能为此做好心理准备，我想你会过得更好。

(01:09:59) Lenny Rachitsky

English:

I need to make that the title of this episode. It's going to get much weirder very soon. I 100% believe that. Oh my God. I don't know what's in store. I love how you're at the center of it all. With that, we reached our very exciting lightning round. I've got five questions for you. Are you ready?

中文翻译:

我得把这句话作为本集节目的标题：“很快事情就会变得更加怪异”。我百分之百相信这一点。天哪，我不知道未来会发生什么。我喜欢你处于这一切中心的感觉。那么，我们进入了非常令人兴奋的闪电轮问答。我有五个问题问你，准备好了吗?

(01:10:14) Benjamin Mann

English:

Yeah, let's do it.

中文翻译:

准备好了，开始吧。

(01:10:16) Lenny Rachitsky

English:

What are two or three books that you find yourself recommending most to other people?

中文翻译:

你最常向别人推荐的两三本书是什么?

(01:10:20) Benjamin Mann

English:

The first one I mentioned before, Replacing Guilt by Nate Soares. Love that one. The second one is Good Strategy Bad Strategy by Richard Rumelt. Just thinking about in a very clear way, how do you build product? It's one of the best strategy books I've read and strategy is a hard word to even think about in many ways. And then the last one is The Alignment Problem by Brian Christian. Just really thoughtfully goes through what is this problem that we care about that we're trying to solve here? What are the stakes in a version that's more updated and easier to read and digest than superintelligence?

中文翻译:

第一本是我之前提到的，Nate Soares 的《取代内疚》。非常喜欢。第二本是 Richard Rumelt 的《好战略，坏战略》（Good Strategy Bad Strategy）。它以非常清晰的方式思考如何打造产品。这是我读过的最好的战略书籍

之一，在很多方面，“战略”甚至是一个很难思考的词。最后一本是 Brian Christian 的《对齐问题》（The Alignment Problem）。它非常周全地阐述了我们关心的、试图解决的问题到底是什么？它的利害关系是什么？相比《超级智能》，这个版本更现代，也更容易阅读和消化。

(01:10:58) Lenny Rachitsky

English:

I've got Good Strategy, Bad Strategy right behind me. I think I'm going to point to it. There it is.

中文翻译:

《好战略，坏战略》就在我身后。我想指给你看，就在那儿。

(01:11:02) Benjamin Mann

English:

Nice.

中文翻译:

不错。

(01:11:03) Lenny Rachitsky

English:

I've had Richard Rumelt on the podcast in case anyone wants to hear from him directly. Next question, do you have a favorite recent movie or TV show you've really enjoyed?

中文翻译:

我曾请 Richard Rumelt 上过播客，如果有人想直接听他的见解可以去听。下一个问题，你最近有特别喜欢的电影或电视剧吗？

(01:11:10) Benjamin Mann

English:

Pantheon was really good based on Ken Liu or Ted Chiang's story. Ken Liu I think. Super good talks about what does it mean if we have uploaded intelligences and what are their moral and ethical exigencies. Ted Lasso, which is supposedly about soccer, but actually it's about human relationships and how people get along and just super heartwarming and funny. And then this isn't really a TV show, but Kurzgesagt is my favorite YouTube channel and goes through random science and social problems and is just super well done and super well-made. Love watching that.

中文翻译:

《万神殿》（Pantheon）非常棒，是根据刘宇昆（Ken Liu）的小说改编的。非常精彩，探讨了如果我们拥有“上传智能”意味着什么，以及他们的道德和伦理需求。《泰德·拉索》（Ted Lasso），表面上是关于足球的，但实际上是关于人际关系和人与人之间如何相处的，非常温馨且有趣。然后，这不完全是电视节目，但 Kurzgesagt 是我最喜欢的 YouTube 频道，它讲解各种科学和社会问题，制作精良。我非常喜欢看。

(01:11:53) Lenny Rachitsky

English:

Wow. Haven't heard of that as you were talking, I feel like Ted Lasso, I feel like that's what you need to put into constitutional AI, act like Ted Lasso.

中文翻译:

哇，没听说过那个。当你说话时，我觉得《泰德·拉索》就是你需要放入宪法级 AI 的东西——表现得像泰德·拉索一样。

(01:12:00) Benjamin Mann

English:

Yes.

中文翻译:

是的。

(01:12:00) Lenny Rachitsky

English:

Kind. Smart-

中文翻译:

善良、聪明——

(01:12:03) Benjamin Mann

English:

Exactly.

中文翻译:

没错。

(01:12:03) Lenny Rachitsky

English:

... Hardworking. Oh my God. There we go. I think we've solved alignment problems right here. Get those writers on this, ASAP. Okay, two more questions. Do you have a favorite life motto that you often come back to in work or in life?

中文翻译:

……勤奋。天哪，就是它了。我觉得我们就在这儿解决了对齐问题。快去请那些编剧来，越快越好。好了，最后两个问题。你在工作或生活中有什么经常回想的座右铭吗？

(01:12:15) Benjamin Mann

English:

Well, a really dumb one is, have you tried asking Claude? And this is getting more and more common where recently I asked a coworker like, "Hey, who's working on X?" And they were like, "Let me Claude that for you." And then they sent me the link to the thing afterwards and I was like, "Oh yeah, thanks. That's great." But maybe more of a philosophical one I would say, everything is hard. Just to remind ourselves that things that feel like they're supposed to be easy, it's okay to not be easy and sometimes you just have to push through anyway.

中文翻译:

嗯，一个挺傻的是：“你试过问 Claude 吗？”这变得越来越普遍了，最近我问一个同事：“嘿，谁在负责 X 项目？”他们说：“让我帮你 Claude 一下。”然后他们把链接发给了我，我说：“噢，谢了，太棒了。”但如果说更有哲理一点的，我会说：“万事皆难。”这只是为了提醒我们自己，那些感觉应该很容易的事情，如果不容易也没关系，有时你无论如何都得挺过去。

(01:12:49) Lenny Rachitsky

English:

And rest in motion while you're doing that.

中文翻译:

而且在做的过程中保持“动态休息”。

(01:12:51) Benjamin Mann

English:

Yeah.

中文翻译:

是的。

(01:12:51) Lenny Rachitsky

English:

Final question. I don't know if you want people to know this, but I was browsing through your Medium posts and you have a post called Five Tips to Poop like a Champion. I'd love it. Can you share one tip to poop like a champion if you remember your tips?

中文翻译:

最后一个问题。我不知道你是否想让大家知道，但我浏览了你的 Medium 文章，你有一篇叫《像冠军一样排便的五个技巧》。我很喜欢。如果你还记得的话，能分享一个“像冠军一样排便”的技巧吗？

(01:13:06) Benjamin Mann

English:

I of course do. It's actually my most popular Medium posts.

中文翻译:

我当然记得。那实际上是我在 Medium 上最受欢迎的文章。

(01:13:12) Lenny Rachitsky

English:

Okay, great. I can see that. It's a great title.

中文翻译:

太好了，我能理解。那是个很棒的标题。

(01:13:15) Benjamin Mann

English:

I think maybe my biggest tip would be use a bidet. It's amazing. It's life-changing. It's so good. Some people are kind of freaked out by it. It's the standard in countries like Japan and I think it's just more civilized. And in 10 or 20 years people would be like, how could you not use that?

中文翻译:

我想我最大的建议可能是：使用洗屁屁机（bidet，净身器）。它太神奇了，能改变生活。它非常好。有些人可能觉得有点奇怪，但在日本等国家这是标准配置，我认为这更文明。在 10 年或 20 年后，人们会说：“你怎么能不用那个呢？”

(01:13:37) Lenny Rachitsky

English:

And a bidet could be like a Japanese toilet. That's along the same lines.

中文翻译:

洗屁屁机就像日本马桶，是一回事。

(01:13:40) Benjamin Mann

English:

Yeah.

中文翻译:

是的。

(01:13:40) Lenny Rachitsky

English:

Right. Okay. I love where we went with this. Ben, this was incredible. Thank you so much for doing this. Thank you so much for sharing so much real talk. Two final questions. Where can folks find you online if

they want to reach out, maybe go work at Anthropic and how can listeners be useful to you?

中文翻译:

好吧，我喜欢这个话题的走向。本，这太精彩了。非常感谢你参加节目。非常感谢你分享了这么多真知灼见。最后两个问题：如果大家想联系你，或者想去 Anthropic 工作，可以在哪里找到你？听众能为你做些什么？

(01:13:52) Benjamin Mann

English:

You can find me online at benjmann.net and on our website, we have a great careers page that we're working on making a little bit easier to access and figure out, but definitely point Claude at it and it can help you figure out what could be interesting for you. And how can listeners be useful to me? I think safety pill yourself, that's the number one thing and spread it to your network. I think. Like I said, there are very few people working on this and it's so important. Yeah, think hard about it and try to look at it.

中文翻译:

你可以在 benjmann.net 找到我。在我们的官网上，我们有一个很棒的招聘页面，我们正在努力让它更容易访问和理解，但你绝对可以用 Claude 来分析它，它可以帮你找出你可能感兴趣的职位。听众能为我做些什么？我认为是“让自己接受 AI 安全理念”（safety pill yourself），这是最重要的，并将其传播给你的社交圈。就像我说的，研究这个的人很少，但它又如此重要。是的，深入思考并关注它。

(01:14:28) Lenny Rachitsky

English:

Thanks for spreading the gospel, Ben, thank you so much for being here.

中文翻译:

感谢你传播这些理念，本，非常感谢你能来。

(01:14:31) Benjamin Mann

English:

Thanks so much, Lenny.

中文翻译:

非常感谢，Lenny。

(01:14:32) Lenny Rachitsky

English:

Bye everyone. Thank you so much for listening. If you found this valuable, you can subscribe to the show on Apple Podcasts, Spotify, or your favorite podcast app. Also, please consider giving us a rating or leaving a review as that really helps other listeners find the podcast. You can find all past episodes or learn more about the show at lennyspodcast.com. See you in the next episode.

中文翻译:

大家再见。非常感谢收听。如果你觉得这期节目有价值，可以在 Apple Podcasts、Spotify 或你喜欢的播客应用中订阅本节目。此外，请考虑给我们评分或留下评论，因为这能真正帮助其他听众发现这个播客。你可以在 lennyspodcast.com 找到所有往期节目或了解更多信息。下期节目见。