Forethought is Hiring Researchers (with Mia Taylor) – Transcript

About this transcript

This is an AI-generated transcript of an interview with Mia Taylor, an episode of ForeCast.

Because it is AI-generated, it may contain substantial errors. You are welcome to cite this document, but consider manually checking the text accurately reflects the audio.

Chapters

Forethought hiring pitch and researcher roles

Fin Moorhouse: Hey, listeners. This is a bit of a bonus episode to say that Forethought is hiring. So if you think you might ever consider applying or know someone who could, then this one's for you. I'll start by just pitching the researcher roles we're hiring for, and then after that, I'm going to speak to Mia Taylor, who is a research fellow at Forethought, a colleague of mine, and we'll speak a bit about what it's like to work at Forethought. Why you might consider applying, stuff like that.

Okay, so in brief, Forethought is looking to hire researchers, ideally either with a strong track record already, or a commitment to learning rapidly. And other than perhaps the importance of the research, you might consider joining for the freedom to focus on the kind of questions you think are most important. There is, I think, fairly good support for turning those ideas into action. There's also, I'd say, a real earnest collegiate sense of discussing big ideas and learning from outside experts.

A few more details. Forethought is looking to hire for two roles in particular. First being senior research fellows to mostly lead their own research directions, and the second being non-senior research fellows who are more developing their own research directions, and will spend more time collaborating with and supporting other research staff.

A few words about senior research fellows. So as I say, they'll mostly lead their own projects, normally in areas which are currently poorly mapped out, and they might also lead a small team of researchers too. And then secondly, non-senior research fellows. Here the paradigm hire might be someone who's perhaps more early career or in the process of switching fields, someone who's in the process of developing their own independent views and research directions.

What would make you a good fit for either role? It would be pretty good if you're already quite invested in some of these big picture questions related to AI progress, and you could take a look at Forethought's research to get a sense of them. You ideally have takes, like I said, and views on what are the most promising research directions that you want to follow yourself or you want to see followed. You have some track record of producing original research. Maybe you just really enjoy getting stuck into a research project and seeing it through to completion. Hopefully you can communicate fairly clearly in writing because that's a big part of the work.

I'll also say we are open to hiring visiting researchers for people who can join for three to 12 months or so. For more details like logistics and comp and locations, there is a job posting on the website, and I'll link to that in the show notes. Applications close on the 25th of October 2025. That's about two weeks from when I'm publishing this, hopefully. And just lastly, I'll say that there is a referral bonus, and it's quite handsome. So if you know someone who might consider applying, who might not be on our radar yet, there is a referral form linked on that page.

Okay, like I said, you're now going to hear a short conversation with Mia Taylor, my colleague, a research fellow at Forethought. We're going to talk about what it's like working here, why Mia herself joined, and why you might or might not think about applying yourself. Okay, Mia, thanks for joining me.

Mia's role and reasons for joining

Mia Taylor: Yeah, great to be here.

Fin Moorhouse: Just tell me, what's your role at Forethought?

Mia Taylor: So I'm a research fellow at Forethought.

Fin Moorhouse: And when did you join?

Mia Taylor: I joined about five weeks ago.

Fin Moorhouse: Great. Why did you join?

Mia Taylor: Yeah, I read a lot of the work that Forethought had been doing over the first six months or so that it had existed, and I thought that Forethought was addressing a bunch of really interesting questions that I didn't see anyone in the space working on. And I think it was addressing a lot of the questions that I felt most confused and kind of uneasy about when thinking about how to make AI go well. I think I didn't have a very clear idea of—suppose we make a lot of technical progress on the alignment problem and we manage to fix all of the issues that come from humans not being able to load some values into AI or get AI to be corrigible or have some other nice property like that—how does that actually cause the future to go well?

And it seemed like there were a bunch of other issues, like concentration of power or, well, whose values get loaded into the AI? Or how do we transition from our state to some really great future that I didn't have a very clear roadmap for, and I was excited to see that Forethought was thinking about that.

Fin Moorhouse: I wonder if a hesitancy a lot of people would have is these sound like interesting questions, but they also sound like questions we can afford to punt.

Mia Taylor: I think that for things like concentration of power, it's not clear whether we can in fact muddle through on that. If power gets concentrated and there are people who used to be in power but are now out of power — which is a pretty simple and stylized way of thinking about it — then it's unclear how we would go and correct that after the fact. I think there is some work like that that does actually need to be done in tandem with developing powerful AI systems. For transitioning to an extremely good future, I'm less sold that there's a lot of stuff that can be done positively in that direction, as opposed to trying to prevent different ways we could fail to get a really good future. But I also think that there hasn't been a ton of thought on this yet, and it seems worth investing in a few more FTEs to figure out if there is anything there because it could be pretty good.

Fin Moorhouse: I wonder, when you were considering whether to join, do you remember what felt like the important considerations — pros and cons?

Mia Taylor: I think the thing that felt like a pretty strong pro to me was that I had a strong sense that I didn't really know what the most important thing for me to be doing was. There are some things I could imagine doing, and I'm not even sure whether they're net good or net bad. It seemed like forethought was pretty open and excited about me doing a lot of thinking and trying to get a clearer vision there. That seemed, in some sense, like the project of forethought.

Fin Moorhouse: So you've been at Forethought for really not very long, but you sit down at the beginning of the workday. What does a day look like?

Mia Taylor: Yeah, so far it's mostly been trying to flesh out what my own views on a bunch of Forethought-adjacent topics are. One thing I've been thinking about is preserving democracy through the intelligence explosion. Just trying to visualize what that would look like. What

would it look like for democratic backsliding that falls short of a coup, for instance, to happen during the intelligence explosion? I don't know if I've come up with anything super original here, but it's mostly trying to lay out my views as concretely as possible and then circulate them with other people who've thought about it a lot and get their feedback on whether I'm missing important considerations.

Fin Moorhouse: Yeah, I wonder if this was the kind of work you were imagining before you joined.

Mia Taylor: Yeah, I definitely think I've spent more time following my own curiosities and confusions than I expected to. I expect that to change over the next few weeks — to move out of this mode of exploring and fleshing out my own views. But I have spent longer in this mode than I expected.

Daily work, examples, fit, cautions

Fin Moorhouse: I didn't know when the last time you were in academia was, but do you have a sense of how this kind of work compares to the most relevant kind of post-grad work or something like that?

Mia Taylor: Interesting. Yeah, the closest thing within academia that I've done is maybe a CS research internship with a professor in my CS department. That felt fairly different — I was addressing a narrowly scoped problem.

Fin Moorhouse: Yeah, I mean, I guess that has pretty obvious downsides, which is that there's not really a path laid out ahead of you.

Mia Taylor: Yeah, I definitely think that's true. It's hard to know when you're starting a project or a sprint and you're trying to get all the implicit parts of your worldview onto paper, then see whether they fit together and whether anything new comes out of that exercise.

Fin Moorhouse: Yeah, I wonder how doomy or optimistic you feel in general about this process of taking one of these really big-picture questions about society after superintelligent AI. Do you have examples of the kind of work that has succeeded so far or feels like it succeeded?

Mia Taylor: I don't know if this has gotten all the way to success for us to count it, but I think the coup project, or the power-grab project at forethought, maybe came from a similar process of trying to think through how things might play out and then realizing there's something really bad that could happen, and then coming up with concrete countermeasures. For most of the questions we're trying to understand, successes — where we produce a reasonably robust, clean recommendation — will probably be rare. But I think there probably are additional successes.

Fin Moorhouse: I remember speaking to Lucas VinVaden, who was a co-author on the coup article, maybe a year or just more ago. And he said, oh yeah, I'm working on AI-enabled coups. I had, I'll admit, a reaction: that sounds kind of silly, but good luck. And I think I was shown wrong in that case. It's not only that I think they made research progress, but also the kind of progress where now I think it's possible to say this is a real worry and that you can make sensible and substantial claims about it. Moving from that sounding kind of strange or silly to, oh, okay, it's good to see that's possible.

Mia Taylor: And I think that—probably I know the history here less well or I've read fewer of the foundational documents—but my sense is that the early AI safety movement was also sort of thinking about how things might play out, being like, oh, shit, this seems like an issue. And then I guess it's still controversial how correct or solid the recommendations were. But at least this was doing a bunch of first-principles thinking. Ten years later, there was decent empirical research on that kind of phenomenon and safety measures being put in place.

Fin Moorhouse: Yeah, I totally agree with that. We can now talk about there being a science of alignment without sounding ridiculous. Maybe one question is, if someone's listening to this, there's a decent chance they are considering applying or know someone who could apply — why shouldn't they apply?

Mia Taylor: My sense is that the people on Forethought's team are mostly people who are generalists, and what we're doing is being curious, hopefully reasonably smart generalists who are trying to gain more and more understanding of a space and then reaching out to experts when we need expert input. It seems pretty plausible to me that if you have really great domain expertise in a specific area—like if you're really knowledgeable about compute governance mechanisms and you can design good governance policies—I think we would love to have someone like that, but plausibly it's not their comparative advantage.

Fin Moorhouse: Yeah, that seems right. Or maybe the way I'd say that back is you are allowed to know a lot about particular things, but maybe if you are the onship verification mechanism person, and that's your entire career and domain of interest, and anything outside of that domain you're like, "Man, not my wheelhouse," then maybe this is not the thing.

Mia Taylor: Yeah, I was thinking more of a reason for you, even if we would love to have you, maybe a reason for you not to join us.

Fin Moorhouse: Yeah. Like our loss, but better for the world.

Mia Taylor: Yeah, exactly. Maybe just thinking about things: this job seems to have a lot of writing and a lot of trying to make your ideas clear and crisp in writing. People might vary in how rewarding it is for them to be writing docs all day. I don't know. I think it is all pretty speculative. And so it seems super reasonable to want something where the path to impact is a lot clearer and a lot more concrete. I think I'm betting on this "space hit impact" model where

maybe most of my work won't end up being particularly valuable, but maybe I'll hit on one idea or one threat model that is really valuable.

Fin Moorhouse: Yeah, that's a really good point. You need to be comfortable with how speculative the work often is, but not always is. If your source of motivation is the flywheel of quickly validating that you did great work and you optimize the software by 10% this month, that's kind of harder to come by.

Mia Taylor: Yeah. Before I joined Forethought, most of my research was empirical. In empirical research, you do have more of a signal of, "Oh, did I find something interesting?" I mean, there's still a lot of interpretation of results and people can vary in how surprising or interesting they find them, but you can show your results to a bunch of people and they'll be like, "Oh yeah, this seems pretty surprising." Whereas I think you get way fewer signals from reality that you're onto something here.

Fin Moorhouse: Maybe a last question. Do you have any interesting disagreements with the rest of Forethought or your impression of what the rest of Forethought believes?

Mia Taylor: Yeah, my guess is that I am more worried about extinction risk from AI than the typical Forethought person. I haven't really drilled into what the cruxes are, so I don't know if I have anything more interesting to say about that.

Fin Moorhouse: Well, that's a notoriously easy question to reach agreement on.

Mia Taylor: Yes, yes.

Fin Moorhouse: I'm sure we can do that at some point. Okay, any other things you want to say?

Mia Taylor: No, I don't think so.

Fin Moorhouse: Great. Me and Taylor, thanks very much.

Mia Taylor: Thanks.

Fin Moorhouse: Okay, me again, thanks for listening. Like I said, the link to the job ad and to the referral form will be in the show notes, but I can also just read it out. It's forethought.org/careers/researcher. Okay, see you next time.