SDS PODCAST EPISODE 791: REINFORCEMENT LEARNING FROM HUMAN FEEDBACK (RLHF), WITH DR. NATHAN LAMBERT

Jon Krohn:	00:00	This is episode number 791 with Dr. Nathan Lambert, Research Scientist at the Allen Institute for AI. Today's episode is brought to you by AWS Cloud Computing Services, and by Crawlbase, the ultimate data-crawling platform.
	00:17	Welcome to the Super Data Science podcast, the most-listened-to podcast in the data science industry. Each week, we bring you inspiring people and ideas to help you build a successful career in data science. I'm your host, Jon Krohn. Thanks for joining me today. And now, let's make the complex simple.
	00:48	Welcome back to the Super Data Science podcast. I'm super excited to have Dr. Nathan Lambert as our guest on the show today. Nathan is a Research Scientist at the Allen Institute for AI in Seattle, where he's focused on fine-tuning large-language models based on human preferences and advocating for open-source AI. He's renowned for his technical newsletter on AI called Interconnects. He previously helped build an RLHF, that's reinforcement learning from human feedback, research team at Hugging Face, and he holds a PhD from Berkeley in which he focused on reinforcement learning and robotics, and during which he worked at both Meta AI and Google DeepMind.
	01:23	Today's episode will probably appeal most to hands-on practitioners like data scientists and machine learning engineers, but anyone who'd like to hear from a talented communicator who works at the cutting edge of AI research may learn a lot by tuning in.
	01:35	In today's episode, Nathan details what RLHF is and how its roots can be traced back to ancient philosophy and modern economics, why RLHF is the most popular technique for fine-tuning LLMs, popular alternatives to RLHF such as RLAIF and direct distilled preference

optimization, limitations of RLHF, and why he considers AI to often be more alchemy than science. All right, you ready for this fantastic episode? Let's go.

02:08 Nathan, welcome to the Super Data Science podcast. I'm excited to have you here. Where in the world are you calling in from?

Nathan Lambert: 02:14 I'm in Oakland, California for a few more weeks, and thanks for having me, Jon.

Jon Krohn: 02:18 Nice, my pleasure. So let's dig right into the technical stuff here. You're currently a research scientist for the nonprofit Allen Institute for AI where you work on reinforcement learning for human feedback, RLHF, and fine-tuning LLMs in the open and for the common good.

02:37 Earlier this year, you released the LLM OLMo, and OLMo has been praised for its openness, including providing pre-training data and training code. So people are probably aware that for the most part, when there's supposedly open-source models like the Llama series or Gemma, they're not open like yours is.

Nathan Lambert: 02:56 Companies are getting better at talking about it. They release internal documents telling people how to communicate about it, but some people mess it up still.

Jon Krohn: 03:06 Yeah, yeah. Well, we really appreciate all the openness. What do you think are the most... What are the reasons driving you behind having as much openness as possible with these AI development tools that you create?

Nathan Lambert: 03:20 It's various. It's wanting to have "a good outcome," which is very biased by what I think of as good, but it's just having a lot of people being able to work in AI, having AI be understood because it's going to be very powerful over these next few decades, and then the worry of probably

largest corporate capture where if AI is as powerful as people think it could, it could result in companies that are 10 times as big as Apple and Microsoft, and then we don't really know how the modern economic system would work in that context.

03:53 So just spreading the love, making sure there's not risks through obscurity and people not knowing what's going on, but then also just education and more people getting involved in these very long-term societal shifts.

Jon Krohn: 04:09 Yeah, on that note of things happening at the Big Tech companies, in your Interconnects.ai newsletter, you recently published an article about OpenAI's Model Spec. So this document details how they steer their model with RLHF toward their goal model behaviors, and yeah, can you elaborate on, I guess, RLHF a bit in general for our listeners that maybe don't know it, and yeah, what your findings were about this Model Spec document?

Nathan Lambert: 04:36 Yeah. So RLHF, this reinforcement learning from human feedback, is the most popular fine-tuning technique right now. It's interesting to people deep in the weeds of language models because it's a different loss function. I think all this pre-training is done and you hear about instruction fine-tuning or supervised fine-tuning. That's with the same autoregressive logs function that's at the core of modern NLP, and it wasn't always the core of NLP, which is fun to learn about, but this RLHF process brings in this human factor, so it lets the models have some humanness that is hard to capture in data. And then it also, it's a really broad loss function. So there's a lot of different things you could try.

05:14 We're just still barely starting at understanding, now that we could do policy gradient updates or different types of updates to these models, how different can we make them

and in very useful ways? And that's a tool. RLHF is a tool for doing this.

05:29 It seems very likely to stick around. If you had had me on a year ago, I would be like, "Oh, I don't know. We're going to try." But at this point, everyone's investing in it. Scale AI did a Series F round on tons of revenue. That's built on RLHF. They've pivoted their company for a second time. And this Model Spec thing is a nascent corporate direction thing, which I think a lot more companies will do.

05:54 The OpenAI, again, is at the forefront of this. It seems like it's a combo work from Jon Schulman who's this author of this Proximal Policy Optimization paper and one of the leading authors of deep RL as a field, before language models come in, and then also their product team. And it's essentially trying to say what we want our language models to do, whether or not we get the technical details right. And then they're going to update this document over time as they better understand their customers and company culture.

06:22 And as a standalone document, it's interesting to people very in the weeds like me, what are the examples they give? What is the commentary they give? They talk about NSFW content and how it's hard to thread that needle and the order of command. So OpenAI has final say on what the model can do relative to the customer and all these things. And just having it in one place is good because eventually, we're going to have multiple model providers do stuff like this.

06:45 And then if you're shopping around, you could see, oh, what does Google want their models to do, or Anthropic, OpenAI? And it's even something that we want to try to do at the Allen Institute just to be like, "What are we trying to do?" We're pretty behind on terms of what the RLHF

practices are. We don't have as big of a budget for human data, but we want to be able to say, "These are our goals and we're going to see if we can achieve them," and then we can document how our goals change over time. And it's just a nice feedback loop to be like, "RLHF is a messy process, but these are the sort of things that seem tractable to do with it, whether or not it actually works."

Jon Krohn: 07:20 Yeah, yeah. And so while AI2, well, at AI2, which is the abbreviation for the Allen Institute for AI, so while you may be working at those details around RLHF still, you've published a popular paper called "Zephyr: Direct Distillation of Language Model Alignment" about how you leverage distilled Direct Preference Optimization, so dDPO, for intent alignment in smaller models, which allows this 7 billion parameter Zephyr model to surpass the Llama 2 70 billion one. So 10 times as many parameters. How does this approach differ from traditional methods and what are the implications for scalability, transparency, accessibility?

Nathan Lambert: 08:02 Yeah, so this Zephyr paper is mostly about the Direct Preference Optimization (DPO) paper, making it mainstream. It's funny, just two days ago, Chris Manning had me for a lecture in his class and he's like, "Oh my god, thank you guys for making DPO seem real." It's like because there was this big time lag. The DPO paper came out in about June of 2023, and then the Zephyr model was released in September of 2023 as the first real model to make a breakthrough with this DPO method, which is a long time when there's so many labs invested in training these models and releasing them for PR and product gain.

08:37 So there's a huge time lag there, and it really built on just strange exploration in terms of experimental details. We needed a really low learning rate. There's the meme in AI that 3e-4 is the only learning rate that you need and it

works for everything, but this model and the model later at AI2, which was the Tulu 2 model, which was 70 billion parameters showing it could scale, both used a 5e-7 learning rate, which is just so outside the realm of normal for most people doing fine-tuning and anything with AI at the time.

09:11 And then there's also this idea of synthetic data, which is this direct distillation idea. So there is a dataset called UltraFeedback from a group that's like OpenBMB, which I think is a research group based in China. And that dataset, and still to this day, if you compare to other preference datasets like Stack Exchange or Stanford Human Preferences or Anthropic's HH RLHF dataset, it's just using the methods that we have on this UltraFeedback data, which is a mix of completions from models like GPT-4, GPT-3.5, Llama 2, and then the chosen and rejected completions for creating "preferences" are labeled by GPT-4.

09:52 So this dataset is six months on from Zephyr or Tulu is still what we're seeing as the best. By the time this is aired, we'll release some more models trained with PPO, or soon after this airs, and the best results still come from using this dataset.

10:06 So as this conversation goes on, I'll probably keep beating my drum, which is we need to get people making more datasets in the open if we want to keep doing this academically and for open source, but I do that in a lot of channels, but it's obvious the field will move so much. We have these Llama 3 models. I'm sure Mistral will come out with something soon, but we're still using the same one dataset. It was the first one of its class too. It's like people don't get the first one perfect, especially when it's a research lab, but it's impressive, the longevity of it.

Jon Krohn:	10:36	Yeah. In addition to Zephyr, while at AI2, you also released Tulu 2, if I'm pronouncing that correctly. There's an umlaut over the first U.
Nathan Lambert:	10:45	Yeah, it's a hybrid camel. It's what a Tulu is. It took me so long to learn this. There's two types of camels and the Tulu is what happens if you cross-breed them. I didn't come up with the name, but that's what it's named after.
Jon Krohn:	10:59	So they're the mule equivalent for camels?
Nathan Lambert:	11:01	I think so.
Jon Krohn:	11:04	Well, so regardless of the etymology of the term, Tulu 2 is a suite of models for adapting pre-trained language models to downstream tasks and user preferences. How does Tulu adapt where other instruction tuning models and methods fail?
Nathan Lambert:	11:19	Yeah, so this is really, the bulk of this project was about understanding the most important instruction data out there. And then we saw Zephyr come out, so we're like, "Let's apply the Zephyr method on top of it." And the Zephyr method, again, worked on these models, so it was a proof of concept that this UltraFeedback data works and at different scales. So it was really the 70B scale, which again was to counter the DPO haters, which is everyone's like, "Oh, DPO works at 7B. No one's going to really use this if it doesn't scale up." And it was like a month later, we're like, "Oh, look, we did it," and it was just funny.
	11:50	But there's a lot of instruction datasets out there and this becomes very messy, I think, if you're deep in the weeds of instruction tuning. You also see these people that are independent affiliations that are trading all these models and uploading them to Hugging Face and they're like, "We have a million examples in our dataset from this list of 15

different places, plus some weird filtering heuristics." And Tulu is the academic version of that, which is focusing on very specific evaluation metrics and trying to understand particularly some things, like code and reasoning, that are harder to improve with the current models, and doing a rigorous study on how you can... It's a two-stage thing that's wrapped into one result, which is how you can improve these with all the data that we have, and then you have a giant dataset, and then how do you prune it down to maintain the maximum performance?

12:41 And what this looks like in terms of Tulu is grad students running a ton of experiments and just getting really deep into the weeds and there isn't a systematic answer. It's like the next thing is to do automated filtering. So the grad students, hey, [inaudible 00:12:54] away from this project, and they're like, "We can't do this by hand anymore," and they're looking into ways of doing automatic filtering based on embeddings or they're looking at influence functions, which I don't even know what they technically are, but it's a measure of similarity between instructions and trying to use this to automate the process of filtering through the vast amounts of instruction data online.

Jon Krohn: 13:16 Nice. Yeah, really cool, all the things you're doing at AI2. It must be an incredible place to work and it sounds like you moving there is to take even better advantage of that, moving out to Seattle.

Nathan Lambert: 13:29 Yeah, it's good. One of the last holdouts of academia and industry hybrid. It's like every workplace, I know it's like everyone has a job. There's always upsides and downsides, but it's a very unique place as industry research has become more closed, but we don't have the resources of industry research, so we have to be a bit more clever in terms of how we do things.

Jon Krohn:	13:48	Are you stuck between optimizing latency and lowering your inference costs as you build your generative AI applications? Find out why more ML developers are moving toward AWS Trainium and Inferentia to build and serve their large language models. You can save up to 50% on training costs with AWS Trainium chips, and up to 40% on inference costs with AWS, inferentia chips. Trainium and Inferentia will help you achieve higher performance, lower costs and be more sustainable. Check out the links in the show notes to learn more. All right, now back to our show.
	14:22	So prior to AI2, quite a bit because you were also at Hugging Face, but prior to that, you were at UC Berkeley and you focused on the intersection of robotics and machine learning, which for me, is personally super fascinating right now. So there've been some really exciting developments in that space recently. Things like NVIDIA's Project GR00T for humanoid robots using generative AI and reinforcement learning. And there's also the announcement of an MIT spinoff Liquid AI, which plans to revolutionize robotics with liquid neural networks.
Nathan Lambert:	14:56	Oh, I didn't even know they're a robotics company. I saw them, but I didn't know.
Jon Krohn:	15:00	Yeah. So what's exciting for you at this intersection of LLMs and robotics? What's promising there for us?
Nathan Lambert:	15:08	Yeah, I think the place that all of this really started was Google Brain's research team on robotics. They're still doing great things, but they were years ahead to embrace this, which is like, "Let's scale up our data engine, let's train some big models." Then it just worked. That'll clearly continue, and on more and more complex tasks as people invest resources in this data.

15:31 There's a product-market fit issue, which is I don't want to buy a robot. So most of the really nuanced takes come on this product side. I'm pretty confident that the research is going to keep going places. I think this is the two biggest trends in robotics and machine learning in the last few years in my mind, one of them is this scaling up data collection, using some sort of large model. You can get real-world results that work. Google showed this, other places have replicated it, they did this open dataset project, but there's also deep RL's actual success area has narrowed and narrowed down to this simulation for robotics where you have procedurally generated worlds and you simulate for robotics.

16:13 I feel like I should try to do another survey on this, but I wrote a blog post a year ago that I was just listing a whole bunch of places where that has worked. It's like drone flight, locomotion, other things. DeepMind had the nuclear fusion paper and it's like there's all these really wild things that have really narrow-scoped deep RLs helping with.

16:32 So I think that's what most robotics companies will be leveraging is we have our robot farm internally that can collect data, and then that's the question of how do you integrate consumer data, or if you're trying to... Humanoids are hard because at a mechanical level, most of the humanoid robots have such high force that it's hard to have them around humans. I think this is what was maybe used to be wherever Eric Jang works, like 10X Robotics or 1X Robotics, they're trying to make actuators that are lower force so it's safer to have them around humans. I think famously, the Boston Dynamics robot, you can't have humans around it because if its arm is doing a motion and it hits you, you go flying across the room because there's so much force. It's like that's not safe.

17:16 So then there's this weird last-mile consideration, which had me really down on it, and then I was talking to a family friend and he's like, "But teleoperation can save you." So it's like if you want to have humans in your house or robots in your house of any type, it's obvious that they're not going to work for many things, but you could outsource the labor to India. There'll be people that'll happily empty your dishwasher manually instead of the robot failing to do it automatically, which creates a redistribution of labor market which arbitrages costs and stuff, which I actually think would probably work if people got over the privacy concerns. So I skipped something in the middle, which is having robots in your house is probably not going to work for a really long time because of the distribution shift, but I think most people are serious about that. And then it's like if the distribution shift is such a big problem, then that's when you do the teleoperation.

18:10 I respect a lot of people that are joining this field right now because there's a lot of opportunity to grab in the language model space in terms of digital applications and building services. It's like the people that are still doing their fundamental research or the people that go to robotics, it's like y'all are taking the long-term thing. I think Eric Jang specifically was like, "Yeah, this is a 10 to 20-year bet for humanoid robotics," and I was like, "respect for taking the big risk," because it does seem to be going in the right direction and robotics has been... If you take away the stable diffusion moment and the ChatGPT moment, the robotics trend line is just the same. It's just slowly, slowly going up and we're pulling in new things, so it doesn't have as much of a splash factor.

18:52 The splash is from people like Elon marketing it now. Tesla Optimus is probably going to be similar to Autopilot. I don't really think of it as exactly what it' marketed as, but they have a really good team there and

they're building a cool robot and then that mismatch will be managed in some downstream way.

Jon Krohn: 19:11 Nice. Gotcha. So basically, you think that it's going to be some time before we have humanoid robots in our homes doing a lot of regular tasks, so where we're going to see more and more robotics applications are in industry typically.

Nathan Lambert: 19:25 Yeah, so there's three really popular Bay Area robotics and AI startups which are Dexterity, Ambi, and Covariant, and all of them have contracts with companies for various logistics tasks like pick-and-place or unloading a truck or loading a pallet. And they all work really well on this. Amazon does this. Amazon is setting up their fulfillment centers to be robot-first. They build entire fulfillment centers from the ground up to be ready for robots rather than being ready for humans, rather than subbing robots in for where there were humans. So all of this really works. It's like how do you get them to leave the manufacturing line type of thing, which is just so different.

Jon Krohn: 20:05 Yeah, yeah, yeah. And so related potentially, because of how we're now seeing LLMs more frequently in robotics applications, like the GR00T explanation from earlier, and you just mentioned Covariant, they had a really cool one too, their Robotic Foundation Model 1, RFM-1.

Nathan Lambert: 20:22 That stuff is the way to get to human interaction. Having language with your robot is the way. It's just going to take a lot of reliability for somebody to want to buy it. It seems so logical.

Jon Krohn: 20:35 Yeah, it's cool, for sure. And so related to LLMs, in your newsletter, you recently wrote an article about GPT-4o that features significant improvements such as latency and real-time audio-generation. In your opinion, which of

these technical breakthroughs will profoundly impact industries beyond the tech sector? So healthcare, education, finance, perhaps making it even into robotics.

Nathan Lambert: 21:02 What are the two? Oh, I think audio is a thing that people will do. I think it's clear that... I mean, I'm the classic example of people that likes to consume a lot of media but does so about through audio. And we've seen this and it's like people watch so much YouTube and so much TV and so much TikTok and it's like how many people watch TikTok versus read Substack newsletters? It's so low. But that removes a barrier to entry in terms of actually using these language models. And GPT-5o, when it exists, it's probably going to be so good in every language. You could just put this in front of a kid that doesn't speak any English and then he has a perfect tutor for anything. Even though most of the education material is in English, it's just ChatGPT already learned all of that and it just exists and it's like the downstream accessibility to education is just so high.

21:53 I think there's obviously social concerns, but it's also starting younger. They don't need to be able to write coherent questions. There's protections you need to add if kids are going to be using this, but I think you could come up with an infinite list, a growing thing of just how talking to these machines with no latency, especially when kids are so clever. My parents would probably be more thrown off or take longer to adapt to an AI that they can interrupt, but a kid probably figures it out in 60 seconds and then they're just going, and they're not even probably going to talk to it like a normal human. They're probably just going to extract the information from it in some unparsable, weird way.

Jon Krohn: 22:34 Yeah, yeah, yeah. I think you're right. It's like I can't really figure out TikTok still.

Nathan Lambert: 22:40 I just protect myself. I'm too deep in the tech industry to know. It's like all the tech kids are protected from all these apps. It's such a sham.

Jon Krohn: 22:49 So going back to RLHF, which has been a huge focus of yours, not only at the Allen Institute but also at Hugging Face where you were previously, in your paper, "The Alignment Ceiling," you discuss the issue of objective mismatch in RLHF. So what are the challenges in aligning reward models with human preferences in a reproducible manner?

Nathan Lambert: 23:14 Yeah, so this paper is a fun story because it hearkens back to my PhD, the PhD work on model-based RL. The objective mismatch paper was my core paper of my thesis, which is essentially in model-based RL, you're learning a policy and a dynamics model. So it's a bit simpler because the evaluation regime is much more closed. So in classic RL, you have these robot tasks and simulations so the evaluation is much more set. And then you can think of it as the dynamics model that is good for the policy is tuned to the policy, it's not tuned to the real world.

23:51 In RLHF, it's a bit different because you're trying to do a multi-stage process where you have this language model, which is your policy, you have this reward model, which is sort of like your environment, but it's mirroring to what the humans want. So there's this extra leg of trying to match what the humans want, and then you have this reward model, and then you have the policy that's trying to extract information out of it.

24:16 There's a lot of analogies. I've been recently talking about information flow, which is if you have this policy that's getting trained, the reward model is some sort of filter or sieve or gain, you could think of it as many different ways, and you need to tune that. You're putting

information through a black box and you need to tune it to make sure that it matches what humans actually want.

24:40 I think this is part of why the Model Spec is interesting. So doubling down on what the Model Spec partially reveals is that when these big companies are collecting preference datas from humans, they have 10 to 20-page documents on, "Here's what you should prioritize when you're labeling the data." And then the question that we haven't been able to see of these models is what is the mismatch between what they tell you to do in the data and what the final model does?

25:04 So if you tell them in the data, "Prioritize factuality, prioritize conciseness," even if the data has that, does the training process result in a model that does this? And that's the best representation of what we're saying is this alignment ceiling is we don't know if our methods ever could actually be perfectly aligned with what our expectations are because we're doing this all in different modules. And then the really deep... This is less in vogue right now, but in deep RL days, people would be like, "Can't we just do end-to-end learning?" So if we only have one objective, can it learn everything at once?

25:39 And that just doesn't seem to scale as well in realistic engineering environments. I think the only people that we see doing that is Tesla self-driving, and we don't know for sure if they are, but all of the teams at OpenAI, Gemini, Anthropic, they have modules where it's like an RLHF team, a safety team, a pre-training team, and that is where they trade off these things. And those are, if you look at this paper, that's what each of those boundaries are where you're trying to design your optimization in the context of the optimization that other people are doing.

	26:08	So it's just this mathematical thing where it's like you can never get a perfect solution if you're doing multiple optimization problems.
Jon Krohn:	26:18	A big alternative to RLHF is Constitutional AI or reinforcement learning from AI feedback. You've discussed that in other places, papers, talks. Can you elaborate on this idea? I know Anthropic is a big proponent of it, for example. Because it seems like some people might be concerned that if you're using AI to judge AI, where's the real ground truth? You know?
Nathan Lambert:	26:48	Yeah. I think Constitutional AI is one of the most misunderstood techniques and this is mostly because the paper is somewhat confusing.
	26:56	So there's two things, there's two major things in this paper you'll see Constitutional AI and RLAIF, and RLAIF is the idea of using AIs instead of humans to label preferences, which is pretty general. And Constitutional AI is a two-stage process which does some RLAIF and some other instruction-revision stuff, where what they do in this paper is they revise the instructions in their instruction dataset with respect... So the completions, they revise those with respect to a list of principles, that's one thing, and then they redo instruction fine-tuning. And then the other thing is they redo this preference label with the context of a principle.
	27:39	So that's what people normally think about and it's just one more way of adding synthetic data. And I think that they've likely moved well beyond that at this point. The paper's pretty old.
	27:51	Some other sorts of things that you could do in this case is you can do a revision to create a preference data. So if you have a bunch of completions from a language model, you could ask it, "Is this factually correct?" And if it says

"No," you say, "can you fix it?" And if it fixes it, then you have a pairwise preference where the chosen preference is the fixed text and the rejected is the original one that needed to be fixed.

28:13 And this is one of a growing list of examples of ways that you could generate synthetic data of which CAI is the one that Anthropic got a lot of buzz on. It's a buzzy name. It was by far and away the earliest. And it's so misunderstood. I mean, I don't feel like I really understood it until a couple months ago. It's just one of those papers that ends up meaning something different than it actually is, which is fine. It happens.

Jon Krohn: 28:45 Nice. All right, so back to RLHF. Given the inherent subjectivity in human preferences, how do you ensure that the aggregated preferences accurately reflect the desired outcomes for our AI systems' behavior?

Nathan Lambert: 29:00 Yeah, we've been debating this a lot internally recently. Essentially, you could phrase it as is the disagreement among labelers a signal or a bug? And it honestly feels more to me that it is a signal because of how vague and multifaceted these preferences are. We don't know what everyone's doing. There's research on this fine-tuned RLHF where you label multiple pieces of an answer, but at the end of the day, a lot of it's being reduced to a pair. And we have different weights. No matter how much we notice different things, like factuality, conciseness, helpfulness, honesty, these very abstract terms. So if this gets reduced, there's going to be some noise.

29:44 And that's seen in the papers where they all report these agreement numbers with their annotators and it's somewhere between 65 and 75% agreement when the people doing research compare their numbers to the annotators. I don't think that's going away. People see that when they use a language model, it has higher

agreement. I don't have a strong opinion on what that means. It could be one of those things where it's like you're amplifying biases when you keep training with language models because they have less disagreements, they have more agreement, and what that manifests as is the diversity of answers may be going down in terms of what's acceptable. We really don't know.

30:27 This is where I said I was going to beat this drum. This is why it's like we need more clearly labeled and with good metadata, open-preference dataset so we can see. Something that I'm encouraging a lot of people to do now is that we have three really good language models. If you're going to do GPT-4 as a judge to label a dataset, do it with all three of them, do it with Claude, do it with Gemini as well, and then we can see what the disagreement between the language models is. And if we get that at scale, we start to learn a lot more.

Jon Krohn: 31:00 Today's podcast episode is brought to you by Crawlbase, the ultimate data crawling and scraping platform tailored for data scientists, AI developers, and Python developers. For ML and AI, high-quality data are of course essential. With Crawlbase, you get a powerful, user-friendly solution that guarantees seamless integration, lightning-fast performance, and unparalleled reliability. Crawlbase supports your needs with a two-minute integration process, AI-powered efficiency, and 99.99% uptime. Crawlbase also excels in bypassing CAPTCHAs, avoiding IP blocks, and handling proxy failures, making them the go-to solution for all your data needs. Use the special code Super Data Science with no spaces to unlock 10,000 free requests. Visit Crawlbase today and supercharge your data collection process with the best in the business.

31:50 Nice. Very cool. Another really cool thing that you've done related to RLHF is you have traced it back to ancient philosophy and modern economics. So mentioning

Aristotle and the Von Neumann-Morgenstern utility theorem, for example. I don't really know what the VNM utility theorem is. But how do these historical foundations influence current methodologies and what can modern AI research learn from these early theories?

Nathan Lambert: 32:20 Yeah. So this was a fun paper with a few colleagues that I started working with at Berkeley and now we're spread out. This is all based on the fact that RL has very deep multi-field, multidisciplinary history where it goes way back. And then the notion of preference is a very vague thing in economics. And it's like the Von Neumann-Morgenstern theory is a foundational thing that essentially, it's like you can express either all behaviors or all goals as probability and expected value distributions, which essentially lets you do expected value math over preferences. And then it led to a bunch of debates on whether or not preferences actually exist and are tractable in any of these things or if they're actually measurable or not due to the preference shift over time based on context.

33:13 So these are the kinds of things that we take and ask a lot of questions on how this impacts the modern RLHF process. It's things like is the final model's preferences, which we're mapping onto very human terms, is that actually based more on the base model, which is scraped from the internet, than the human preferences that they get from somewhere like Scale AI?

33:36 So if it's based more on the internet crawling than this million-dollar dataset they're getting from Scale AI, it's confusing to the marketing where we're saying we're learning a preference model, but it might not actually do that much. There's other things like OpenAI now has a ton of user data and it's like what does the economics literature say about generating data for training that comes from a user context or a professional context where

someone is paid to do it and they're paid to act in a certain way and how does all of this mix?

34:04 So it's really just a super long list of questions of why we should look at other social sciences if we're making grand claims about human preferences and all of these things.

Jon Krohn: 34:16 Nice. Well, fascinating. Tons to dig into there for our listeners. Final topic that I planned related to RLHF, I'm sure it'll come up again organically in the conversation, but you've mentioned that RLHF is not even robust to fine-tuning. And so removing the safety layer from models, like GPT-4 and Llama 2, can break down the notion of safety. Can you elaborate on the implications of this fragility for the future development and deployment of AI systems?

Nathan Lambert: 34:49 Yeah, so this is a specific line of research. There was a few papers that showed that if you take a model like Zephyr or Tulu that we were mentioning, if they have safety in the dataset, if you then go and fine-tune it again on some different tasks, you'll lose some of the behaviors that are "ingrained" in the model.

35:07 I honestly think this is a little bit more clickbaity than actually worrisome because it's really not surprising, given that if you just look at the amount of compute applied at fine-tuning, we pre-train these models for trillions of tokens and then we apply a couple billion tokens of compute at fine-tuning, and it's like we're not changing the weights of the model substantially. We're doing a slight nudge and it makes sense that a slight nudge could be undone at the same way.

35:34 But if you are to take this to some of the bigger labs, what you hear is that safety is not just a single artifact thing. Safety is much more about a complete system than a model. So open-weight models being safe or unsafe, I

don't consider it to be that big of a deal. It's like if you were to apply them to a free endpoint that everyone on the internet could talk to, then I don't want my model saying good things about Hitler and all these obvious things. But if it's a research artifact that you need to spin up GPUs to use yourself, it's a little bit more... I'm more open to having these diversity of models exist.

36:11 But if you ask Anthropic or somebody, it's like, "What happens if... How do you get safety into your model?" And it's not just RLHF. You need to have safety at the pre-training, any preference model you train, and then all of these models have a safety filter on the output. So ChatGPT, it reads all the texts generated from the base model and then there's go-no-go where it will rephrase the text if it gets a no-go signal, which is their content moderation API.

36:35 So it's like it's a double... It's the type of thing where researchers need to market their work, but it's not as big of a deal as I think it is. It's like, okay, I think it has interesting business downstream things with liability. So it's just like if you want to fine-tune a model, you normally do that on your own hardware, but OpenAI has a fine-tuning API and if they claim their model is safe, but any fine-tuning on their API that they then host makes it unsafe, that seems like more of a business problem, which is like, oh, it's a nice way that the open ecosystem might be better off because it breaks the liability chain, but we'll see this research continue to evolve. It's so early in all of these things. We're a year in.

Jon Krohn: 37:21 Yeah, that is something that I had not thought of is how if you're fine-tuning the OpenAI model via their API, you are potentially removing some of the safety stuff, which hadn't occurred to me. Yeah, so moving on from RLHF into some other topics, in your podcast, The Retort, you've discussed AI being "closer to alchemy than

science." Could you elaborate on this perspective and its implications for how we understand and develop AI tech?

Nathan Lambert: 37:50 Yeah, I think a lot of this is about the culture of AI, which it really does, when you're on the ground, feel like things will just work. And there's a lot of people that... It's like you're operating at a scale where real hypothesis testing doesn't really work. It's like we do 10 experiments at the 7B scale, and then they're like, "We're going to train a 35B parameter model based on some reading the tea leaves and intuitions of what we have seen," because we don't have the infrastructure to do thorough testing, which is proper randomization and really thorough, little things.

38:25 So it's definitionally not that scientific and there's a lot of people in the field where it's just not. It's like science is a very clear that people are taught and it's a lot of just like, "Oh, we're going to try this because it feels right and it probably works." And then there's the whole culture thing of how these companies cast narratives about their things being pseudoreligious artifacts and all the AGI talk and stuff, which makes it unscientific in many ways. I think it's a lot more of my co-host Tom's favorite thing to talk about, but I understand the argument and I agree that it's apt. It's been like this for a long time where it's deep learning being uninterpretable fundamentally makes it hard-to-do science.

Jon Krohn: 39:12 Totally, yeah. It is wild and it's crazy to me how when you say, "Okay, with GPT-4, we're going to have 10 times as many parameters as GPT-3." GPT-5 will probably be that same kind of change. Even the people themselves who are developing these systems don't know what emergent capabilities they'll have, so I think that relates to this idea of it being an alchemy.

Nathan Lambert: 39:37 Yeah. There was this NeurIPS's best paper last year by Rylan Schaeffer who gives fun talks. He's really good at storytelling for his papers, but the best paper essentially said that the emergent properties, a lot of the measurement is through statistical measurement error, which is we have all these benchmarks where the random floor is 25%, so then weird statistical things emerge when you finally get signal.

40:02 So if you're training these things, what the test set loss would look like is probably a line going up on log compute, versus as you get X-axis is log compute and the Y is performance going up linearly or something like this, but if you have a noise floor at 25% or something weird where it's not having any signal and then it kicks in, it looks like instead of being a straight line, it looks like a flat that then goes up.

40:31 That's the whole idea of the paper is that most of these arguments are measurement noise. I think it's probably somewhere between that and reality, which is like we are discovering things that are unpredictable with the largest models, but the way that we're presenting them flatters this emergent hypothesis just because of the way that benchmarks were created. So I think that was interesting. It's one of those papers that really should be a blog post because the idea is so clear, but we have to go through the academic gating cycle, so it ended up being a paper. It's just pretty funny.

Jon Krohn: 41:01 Another topic that you covered in your podcast recently was you discussed the idea that RLHF could be fixing something in the pre-training and that it might be correcting biases from common data sources like Reddit. So what's the problem there with using those kinds of common data sources like Reddit and how is RLHF addressing those biases and implications?

Nathan Lambert: 41:25 There's two things here. I've been parroting this theory that fine-tuning is important even if it's not as much raw compute because of how you present information being so important. I like to use this analogy of the Sapiens book, which is obviously stuff that's in history class, but he rewrote it in a way that was so compelling that is one of the most-selling books of all time. And RLHF is doing that on a small scale, which is all these base models have similar things in them, but the models that really resonate with people happen to be output in a way that is really, really compelling.

41:56 So that's the base case of RLHF is style transfer and still being important to just get this flow of the model right. And then there's other stuff that OpenAI makes a lot of funny noise about. Well, they don't make noise about it, but the leaks do, which it's like all this Q* stuff and adding extra search at the fine-tuning phase, which is various ways of just getting very new types of data.

42:19 And it's what I was talking about at the beginning with this different loss function. It's like the way to exploit the fact that we're no longer doing autoregressive loss and see how far that lets us create different types of language models or other types of ML models, which I need to make a talk on this, which is why people are bullish on RLHF, which I haven't done it. I think I need to learn a lot about it though because it's hard to make it more than two slides. It's like what does it actually mean that you're doing these policy gradient updates rather than this autoregressive loss?

Jon Krohn: 42:48 It's wild to hear someone who is so expert in RLHF describe it like that.

Nathan Lambert: 42:52 Someone at OpenAI gave a talk about something like this. I think it would've been... I don't remember. There's the NYU professor that's also at... Maybe his last name is... I

don't know how to say it but I could find it later. But he had this slide in his talk, which is the language model 101, and that was how he presented RLHF, and I was like, "That's a good way to do it."

Jon Krohn: 43:10 Nice, yeah. Beyond all of your professional work, which we've discussed so far, about a year ago, you wrote an article in your newsletter called, "Behind the curtain: what it feels like to work in AI right now. Fear, FOMO, and the scientific exodus driven by ChatGPT." I totally feel this too. It seems so hard to be able to find terra firma, something constant that you can just be like, "Okay, investing in understanding that, is going to be great for my career for years to come." It seems like everything's moving so quickly, but yeah, it is scary. So yeah, so I don't know if you want to talk more about that article and basically-

Nathan Lambert: 44:01 It's settled down a bit. I think this was a transition period to where we're at now. Sorry to cut you off a little bit, but it's like the pace is just so high, but I think there are fundamentals that you still... Learning how to use language models is good. It's almost like when I started my PhD, it's learning anything to do with deep learning and PyTorch and all these things is good. And I think Hugging Face Transformers is a place to start with things. It's good to play with different models. I think it's an in-vogue thing in the industry to be like, "Oh, their code isn't very good. It's not very optimized," but if you're a first-year grad student, it's easy to play with a ton of models and that's what their business is about. It's enabling people to use this, and those things pay off.

44:46 I play with stupid AI things. I transform my newsletter into AI-generated voice stuff, and just getting used to working with all of these things is now the fundamental skill that will pay off in 10 to 20 years because there still probably will be something like an OpenAI API. It'll just be

much better. And that's like Karpathy's take that language models are a new computer processor type of thing. They're just a fundamental computing unit that is worth getting used to.

45:15 And this article was when we were all readjusting to this. I think it's actually a bit better now. I think there's still a lot of people complaining. I think that it was like yesterday, Yann LeCun tweeted this thing, which is, "If you want to have true impact in AI, don't work on language models." And I just feel like there's so much gatekeeping against telling people, "Just go be excited and try things," which was one of the... I quote-tweeted to say the opposite, which is, "You could have a nice life and work in language models." So it definitely brings out the haters for some reason. They're like, "This is a bad take." It's like why? You just have to go play with things. And if you're building things, it's much less important what the noise is because you're actually doing things rather than sitting back and getting bombarded with these random release this, release that, which is obviously cool, but most of them don't matter and it's just good to get grounded in actually doing stuff.

46:12 I don't know if that answers your question, but I don't expect this to change. This is fundamentally driven by the VC cycle where it's until these companies that get really ridiculous funding rounds start to die, we're going to be in this cycle where there are so many releases all the time because that's what all these startups need to do to get customer inbound, to get PR. It's like until these companies start dying, it's going to be the same. And then after that, it's this aggregation phase and collapse, but we don't know if that's going to be a year or five years from now.

Jon Krohn: 46:41 And I like your point in there that LLMs, they make things, they provide so many different kinds of

applications that we can be building now that we couldn't before. And another great advantage of them is that it also just, it makes my workflows a lot easier, whether it's an inline code assistant or something like Claude, which I use for questions that I just have about the world all the time. So yeah, it is exciting, but it can be scary at times. And so-

Nathan Lambert: 47:12 It's worth reading that article to understand what the zeitgeist was like if you were a student. That was really built on just talking, from everyone at Hugging Face, that's all just was freaking out collectively and I was just like, "I'm just fired up one afternoon. I'm just going to go put all of this on a page." I think it's a good capture of the moment when something goes viral like that. I don't think I could recreate it because we're in much of more of a steady state now. We're no longer in this super high-entropy state. It's noisy, but most people are used to it.

Jon Krohn: 47:45 Yeah, yeah, makes sense. So in terms of being able to have a better work-life balance, I mean, to not be fearful of things, to just allow mental health to always come first, which is something you've written on your website, you've managed to integrate your passion for cooking, fitness, and health into your daily routine. How does this influence your approach to work and do you think this is helpful for achieving work-life balance?

Nathan Lambert: 48:10 Yeah, I think the realistic take on what I do is a kind of contrived, partially what I still self-identify as because it helps me be healthy because I have a long history of endurance sports from college and growing up and I still do some of this stuff, and it's like a basic rule tree of do I sleep well, yes or no? And if the answer is no, it's like I feel [beep] during training. And it's like I just wanted to feel good at these things. And I just lean into this enough to make it obvious that I'll not work late. It's like I don't

need to. But obviously things come up, but I don't think I've ever done an all-nighter at grad school or anything.

48:51 And this is how it works for me is training and trail running and getting outside and it's mostly finding what works for you. And I do think it's worthwhile. That was a part of why I want to be in person. I think a lot of people in CS and related fields are so good at optimizing and organizing their life, which is being remote is so convenient that I can overdo this, which is I'll be like, "Oh, well, it's 5:00 PM. I got to go do my workout now." Where it's like if you have to walk 20 minutes into the office, it's just a bit more structure to take back control of your life, which is why that's going to help me a lot. And you just need to have things in your life that make that the case, to just have some things that you don't control, and things outside of work.

49:41 It's hard though. I think it's still coming out of COVID in that regards. I mean, you do remote interviews a lot of the times. I don't know how to balance that. I've thought about interviews from my blog and I'm considering whether or not I only do them in person, which is I go to a conference and I bring a microphone and I see if I can get somebody interesting, which obviously reduces my throughput, but it's just trying to make these rules in my life that reinforce doing things that are actually in the real world. So I think about it a lot and I don't think there are perfect answers.

Jon Krohn: 50:13 Yeah, it's crazy. For me personally, the pandemic was... I mean, it still continues to have a negative impact on me in a lot of the ways that you're describing because I used to have this routine of going to the office and being around coworkers all day, and that laughing, I really enjoyed being at work, and now I mostly work from a home office. And you talked about over-optimizing there. I picked an apartment that has a CrossFit gym across the

street, but now I regret that decision because I'm like, "I can spend my whole day in my apartment and the only thing that I go to do on a typical day is go down to the gym that's across the street." So I'm not like... Yeah, I've over-optimized and yeah, working some things out there myself.

51:09 Anyway, so thank you so much for this great, rich conversation on robotics research, RLHF in particular. Before I let my guests go, I always ask for a book recommendation. I don't know if you happen to have one for us.

Nathan Lambert: 51:27 I'm reading some things that I like right now. So I'm finishing The Three Body Problem series. So the second and third books, I don't think have the peak level of the first book. The first book has two moments in it that I won't spoil that I think are some of the best sci-fi moments in any literature, so Three Body Problem is really worth reading for these key moments. And then it's just great, hilarious sci-fi, which is good for getting out of these things that we talk about.

51:54 And a timely one for people in AI is I've been reading Going Infinite by Michael Lewis on the Sam Bankman-Fried FTX stuff, and reading this, especially if you read some of this behind-the-curtain article, it's like, we're going to get this about AI. I think there's memes on Twitter where people are saying that the title is going to be Not Consistently Candid and it's going to be about Sam Altman, but I think Sam Altman has achieved real things. His success is not going away. He might just have personal things that make it harder, but there are going to be other books where it's like these AI companies come and go in the most dramatic fashion, and just reading that to make sure that you have a good sniff test of total [beep] is probably good.

	52:37	I could probably dig up a more evergreen and non-classic techie book recommendation, but those are the things that I'm reading and I'm enjoying them and that's normally good enough.
Jon Krohn:	52:49	Yeah, that's great. Great recommendations. Three Body Problem has come up quite a few times as a favorite book on the show, but Going Infinite is new and that sounds great. I love Michael Lewis. I've loved him for years.
Nathan Lambert:	53:03	Yeah. It got weird reviews, so I took a second to not get into it, but it reads really well, like all of his books. So I'm like, "This is just..." It obviously has bias, it's not perfect, but it's solid.
Jon Krohn:	53:15	Yeah, he's unreal at making what could be quite dense topics, like Kahneman and Tversky, into a page-Turner.
Nathan Lambert:	53:30	Yeah.
Jon Krohn:	53:31	Awesome. All right, so in terms of following you after this episode, we've already talked about your newsletter, interconnects.ai, we've talked about your podcast, The Retort. How else should people follow you other than those?
Nathan Lambert:	53:42	Those are really the main things. I'm @NatoLambert on most platforms, but somewhat begrudgingly. I only really use Twitter for random thoughts. I'll promote my work on other channels, but the only other additional point of information is going to be Twitter. But I try to make good things in the blog just for my own sake. But it's good to try to condense them down to be less noise. It's always easy to tweet more, but you don't necessarily gain more from tweeting in terms of actually learning anything.
Jon Krohn:	54:15	All right, Nathan, thank you so much for being on the show today. Such an awesome guest, and yeah, we really

appreciate you taking the time and hopefully we'll catch up with you again in a few years.

Nathan Lambert: 54:24 Yeah, thanks for having me. This was good questions.

Jon Krohn: 54:32 What an eye-opening episode. In it, Nathan filled us in on how dDPO allows for intent alignment in smaller models, allowing Zephyr 7B to surpass Llama 2 70B on some benchmarks. He also talked about how Dexterity, Ambi, and Covariant are the big players in robotics, but that today's humanoid robots have too much force to be around in many everyday situations. He talked about how RLAIF can scale up fine-tuning at a lower cost and help resolve whether disagreement among human labelers is a signal or a bug, and he talked about how RLHF can have a positive social impact by fixing the pre-training biases that crop up due to pre-training LLMs on data sources like Reddit.

55:16 As always, you can get all the show notes including the transcript for this episode, the video recording, any materials mentioned on the show, the URLs for Nathan's social media profiles, as well as my own, at Super Data Science .com/791.

55:28 If you'd like to engage with me in person as opposed to just through social media, next week I will be at the Collision Conference in Toronto. It's a four-Day conference. On the Thursday of the conference, I'll be hosting an afternoon of sessions on the content creator stage. Beyond the sessions that I host, other amazing speakers you can check out include the godfather of AI himself, professor Geoffrey Hinton; we'll also have Aravind Srinivas, the CEO of Perplexity; Aidan Gomez, CEO of Cohere; and the tennis legend, Maria Sharapova.

56:02 Thanks to my colleagues at Nebula for supporting me while I create content like this Super Data Science

episode for you. And thanks of course to Ivana, Mario, Natalie, Serg, Sylvia, Zara, and Kirill on the Super Data Science team for producing another fantastic episode for us today.

56:16 For enabling that super team to create this free podcast for you, I'm so grateful to have the sponsors that we have. You can support show by checking out our sponsor's links, which are in the show notes. It's a huge help to us if you do that. And if you yourself are interested in sponsoring an episode, you can get the details on how by making your way to jonkrohn.com/podcast.

56:37 Otherwise, share this episode with people who would like it, review the episode on whatever platform you listen to it on, subscribe if you aren't already a subscriber, but most importantly, just keep on tuning in. So grateful to have you listening and I hope I can continue to make episodes you love for years and years to come.

56:54 Till next time, keep on rocking it out there and I'm looking forward to enjoying another round of the Super Data Science podcast with you very soon.