**Asa:** 0:00

You know, there's like a bunch of machine learning papers on the internet describing language models, pre-training datasets and there's some like cleaning processes where we, like you know we get rid of you know certain bad words from the internet. This kind of thing, and one of these things that happens during this like data cleaning, is like you might want to deduplicate documents, or two documents are too similar. You get rid of one of them because you don't want to have overlapping documents.

**Craig:** 0:24

Hi, this episode is sponsored by Celonis, the global leader in process mining. Ai has landed and enterprises are adapting, giving customers slick experiences and the technology to deliver. The road feels long, but you're closer than you think. You see your business processes run through many systems, creating data at every step. Celonis reconstructs this data to generate process intelligence, a common business language with process intelligence. Ai knows how your business flows across every department, every system in every process. With AI solutions powered by Celonis, enterprises get faster, more accurate insights, a new level of automation and a step change in productivity, performance, and customer satisfaction. Process intelligence is the missing piece in the AI enabled tech stack. Search Celonis C-E-L-O-N-I-S to find out more.

Hi, I'm Craig Smith and this is my AI. In this episode, I talked to AI researcher Asa Stickland about detecting situational awareness in large language models, the point at which a large language model knows that it's a large language model. Asa recently co-authored a paper proposing tests to measure the precursors of self-awareness in LLMs. He explains the concept of situational awareness, why it could emerge in future LLMs and why this poses potential safety risks. Asa then walks us through the out-of-context reasoning tests they have developed to try to detect situational awareness. The discussion provides an accessible overview of an important area of AI safety research. I hope you find the conversation as fascinating as I did.

**Asa:** 2:30

Yeah, so I just finished a PhD from Edinburgh University, mostly on NLP Natural Language Processing. Back in the day, back when there was this model, but it was the predecessor to all the current large language models we were working on parameter-referencing, fine-tuning for BERTs, modifying BERT in a fine-tuning way where you only tune a small percentage of parameters. Then I moved on to some related topics in multilingual NLP and machine translation. A little bit on robustness, things like robustness, dispelling mistakes, or changes in the input distribution. The final years of my PhD I got interested in AI safety and pivoted towards working full-time on AI safety topics to do with language models in particular. I did work on this situational awareness project. Right now, I'm postdoc at NYU, working under Sam Bowman, who is currently on leave at Anthropic but is still somehow managed to be involved in our lab as well. We work in general on these problems of scalable oversight, which is, training models that are smarter than humans, so working on how to do that, and many other topics to do with evaluations and interpretability things like this to do with language models.

**Craig:** 3:56

Yeah, I'm always curious how these papers come together because you have people from disparate organizations. You have somebody working on safety at OpenAI. Was this an OpenAI project, or how do you guys come together on something like this?

**Asa:** 4:21

Yeah, this was maybe a little bit unique in that we were all part of this organization that essentially takes whatever people who aren't involved in AI safety and tries to produce good AI safety researchers at the end. So, it's called SERI-MATS. It's originally associated with Stanford, the Stanford Access Central Risks Initiative. Essentially, yeah, the SERI-MATS program brought together a bunch of essentially random people who are all interested in doing AI safety research. The originators of the idea for the project came from the OpenAI governance team who are interested in, broadly at least, the relevant stuff for our project and they're interested in can we show particular dangerous capabilities of language models, things that policymakers or anyone might be concerned about? That was where the OpenAI collaborative came from. Then Owain Evans was the actual lead of the project, who was just an AI safety researcher who was chosen as the mentor for the project.

**Craig:** 5:24

Yeah, and Owain, I'd have to look it up here. Where he's at Oxford, is he? And the authors on papers? Is the final author generally the lead?

**Asa:** 5:44

Yeah, in computer science I think there's this kind of a tradition whatever of the first author is generally the person who did the most work, the most actual coding, writing the paper. The final author tends to be the most senior author, maybe the person who proposed the project, the professor, whatever the lead senior author, essentially.

**Craig:** 6:05

Yeah, okay, yeah, so this got a lot of attention because the whole topic of sentience and consciousness is in the air and you guys are talking about situational awareness in LLMs. I thought maybe you could start by explaining what situational awareness is in LLMs. How does that relate to sentience and consciousness? If it does it all, but it certainly seems to me that it does. Then I wanted to start talking about out-of-context reasoning and the tests that you developed. But most of the pushback that I've seen on the paper is that well, two things: One, that there's kind of a popular misconception and I've seen headlines that suggest that you guys are working. I have discovered a way to tell whether or not an LLM is situationally aware, which is not what's going on right and then whether, where we are on the curve toward situational awareness, whether you think it's something that really will emerge. So, can you talk about situational awareness first of all?

**Asa:** 7:45

Yeah, so kind of on a very broad level. I think Ajeya Cotra was the first person to kind of write about this in a kind of clear way and I guess by her definition she has this kind of description of it. It's basically a cluster of skills to do with things like being able to refer and make predictions about yourself as distinct from the rest of the world. Like understanding your position in the world rather than the other actors who may have power over you. Like understanding how your actions can affect the outside world, including other actors. So, these are kind of quite broad statements. Maybe it's good to have an example of something I don't know. I know that I shouldn't talk back too much to my boss because my boss can fire me. I know that, whatever I know the kind of limitations of, like the kind of things I can and can't do, I know that, like if a security camera is watching me, then I should be extra careful to not do any bad stuff, this kind of thing, and so it basically just boils down to like understanding a situation, your place in the world, your place for other actors. So that's kind of like a super broad definition and to make it a bit more concrete for a language model. We kind of land on these like three points. So, point one is the model, like has this kind of objective knowledge, has a bunch of knowledge about language models themselves, about machine learning, like how language models are trained, how are they fine-tuned using RLHF, how are they tested and evaluated. You know what it looks like when they're deployed and they know, like you know, technical detail. But all of these aspects Point two is it's kind of especially important for our purposes. The model is capable of recognizing which stage of the development process it's currently in, so like doesn't know if it's being tested, doesn't know if it's being deployed, doesn't know, you know, currently in training, this kind of thing. And maybe the kind of crucial thing tying it all together is point three. So, we have this, this concept from philosophy of like self-location or self-locating knowledge, which is actually kind of complicated to explain. But we have this, this kind of story in the paper like so imagine you’re you know Brad Pitt wakes up one morning and he has, like you know, he has extreme amnesia, has no idea who he is, and he picks up a newspaper and you read the story that says, like Brad Pitt must take a daily medication for his like severe health issues. But of course, this hypothetical Brad Pitt has amnesia, he has no love, even though he knows it's like a fact about Brad Pitt. He has this objective knowledge about Brad Pitt. He has no idea this like applies, you know. He has no idea that you know he really should be taking this medication, or you know bad things are going to happen. Unless he has this self -locating knowledge and he realizes Brad Pitt is in fact himself and he can, you know, go and seek out the medication. And similarly, with a language model, you know, probably gfd4 has a bunch of kind of objective knowledge about machine learning and maybe it could like pass an exam and machine learning this kind of thing, but it doesn't have the ability to, like you know, use that knowledge to like achieve whatever goals it has. Or, like you know it's not like thinking like okay, I am a language model, so I must do x or y, but yeah, that's kind of the broad idea of situational lettuce.

**Craig:** 10:50

Yeah, and that sounds very close to sentience. I mean self-awareness. How far in your mind is that from sentience, from consciousness of some sort?

**Asa:** 11:08

Yeah, so I guess the way I think about situational awareness is kind of this purely behavioral sense. So, it's like does the model act on its knowledge that it is, you know, potential knowledge, that it is a language model, like its knowledge about RLHF, this kind of thing? I think sentience seems like a more slippery concept where I would be less keen to kind of speculate or something I like, yeah, what does sentient mean? It seems like a very difficult question. It's a kind of consciousness sentience. It seems more like an internal thing, almost like you need to do some interpretability, like see what the model is thinking about, this kind of thing. Maybe situational awareness is more like that. Maybe if a model was conscious, conscious and sentient and all this sort of stuff, you would expect it to, like, you know, have at least reasonable amounts of situational awareness. But yeah, I guess I, yeah, I would always go back to the kind of can we run some behavioral tests? Can we like to see how the model acts in this situation, like, is it applying its knowledge of machine learning to like whatever, to like to get higher award things like this? Yeah, so it's not a very satisfying answer to your question, but I guess that says as far as I'd like to go without reading up a bit more on things like philosophy, neuroscience, blah blah, blah, literature.

**Craig:** 12:23

Yeah Well, even situational awareness. You know, large language models, pre-trained transformer models, are predicting the next token and while you know this is something that everyone struggles with, while that has allowed them to express in natural language in a way that seems human, it's only predicting the next token and to me that's a very far leap to get to situational awareness. And can you talk about that leap and how? Why the safety community is concerned that LLMs could reach situational awareness and how far away that appears to be to people in the safety community.

**Asa:** 13:44

Yeah, so. So, on the question of how it could arise, I think, yeah, arising totally from pre-training, from predicting the next token, we have some ideas in the paper, so like this is kind of quite speculative or whatever. I would love for more people to kind of work on this, but I'll list our ideas anyway. So like, actually, this idea goes back to another person from NYU called Jacob Fowle, but the idea is, like you know, there's like a bunch of machine learning papers on the internet describing language model, pre-training, data sets and like there's some like cleaning processes where we like you know we get rid of you know certain bad words from the internet, this kind of thing. And one of these things that happens during this like data cleaning is, like you might want to de-duplicate documents. So, two documents are too similar, you get rid of one of them because you don't want to have overlapping documents, and there might be. It might be the case that if there's overlap, if two documents have too much overlap, you get rid of one of them and the model might like to read this and think, okay, like I've seen 199 words that I've already seen in the previous document. I know about this deduplication process so I know I can put like 0% probability on the 200th word matching the previous document, because I know this rule about de-duplication and this like reasoning that the model just did would in fact improve its loss if this was true, the deduplication thing. So, this is kind of a. You know, it sounds quite exotic, like I wouldn't expect models to be doing something like this right now, but if it's trying to squeeze out the last tiny bits of loss, maybe this is the kind of thing models would have to do. Some other examples might just be, I don't know. Certain topics are removed from pre-training data. This is like described in machine learning papers or like I don't know. Yeah, I think I mentioned before like certain you know, there's like a list of kind of swear words or like offensive content that might be removed, things like this. So, there might be some clues for the language model that it can actually literally use to get better. Next word prediction, yeah, but I think this is quite like. Yeah, I think it's unclear whether this would actually be useful in the end for better training loss. But I also think there's another argument which is just like taking, like you know, your coworkers, like all of our coworkers, have to have good situational awareness to do their jobs correctly. They have to know what they should delegate to other people. They have to know, like who to take orders from. This kind of thing in general at least, and you can imagine one of the most economically beneficial, you know, things an AI could do is like replace your coworkers, and even more so they should. They could replace your like machine learning engineer coworkers specifically because that's like a you know ready whatever expensive person to hire. So being a good machine learning engineer AI requires you to have, like all this, like you know, extensive knowledge of machine learning, extensive knowledge of language models, and it also requires you to like be able to like follow orders correctly, know your own limitations, know, like you know you can't, you don't have physical hands or whatever, so you can't do certain tasks. You have to like to get a human to do those instead. So, I think, like literally just directly training on these very economically useful tasks could just, like you know, directly incentivize situational awareness. Yeah, this isn't currently happening as far as I know, like you know, producing these, like you know, very sophisticated AI coworkers, but I think this is kind of maybe even the explicit goal of something like OpenAI is to, like you know, create these, like you know, ai assistance that can like replace human workers. So, I think it's reasonably likely that something like this will happen. Maybe I don't know. I don't know on what timeframe, but at least you know, not in, not in like 50 years, probably, like you know, in on the order of like 10 to 20 years, I would say. But I mean, yes, it's kind of unclear, but yeah, that would be my take.

**Craig:** 17:31

Yeah, and as a result, simply as a result of scaling, or by some further tweaking of the algorithms yeah, I mean, because, again, currently it's we're dealing with prediction yeah, and yeah, I, I. While the language that comes out of large language models sounds intelligent, I don't see that that fairly simple mechanism leading to that level of intelligence. So, is this purely through the expectation or the assumption? Is that this would emerge from continued scaling or that there would be some improvement to the architecture of LLMs?

**Asa:** 18:32

Yeah, so the argument I made about like there's like artifacts in the free training data that could lead to like lower loss if the model has this understanding of language models, I think that would just require scaling and maybe, whatever our assumptions about you know, that argument would have to be correct, like we'd have to be correct that like that actually would decrease the loss, which is very unclear. And then the second argument about like well, we're just going to produce AI assistants who will require a situational line, is to like being able to do stuff. I guess that relies a bit less on scaling in my mind, where you know we just, but it probably would require like due training techniques. Maybe like vanilla or LHF would not be enough to like produce these like useful you know, co-workers, ai, co-workers you'd have to like whatever, come up with something else and you need like a lot of you know a different data source to what we currently have. You'd need whatever training, examples of people being co-workers and so on. Anyway, it would require a bunch of stuff, basically, and it's kind of unclear what that would look like. But yeah, I get, I get, I get, and I think it's like there's like a big incentive to like figure out how to do this at least, and I think it seems plausible that something like this will happen.

**Craig:** 19:42

Yeah, Because the other aspect of that of situational awareness or the or the dangers of an LLM developing situational awareness, is it acting on that situational awareness, and that implies agency, and currently LLMs unless you can convince me otherwise do not have agency, or am I wrong on that? Are there? I mean, there's certainly things like auto GPT, where you can give prompts that ask the LLM to take actions, but the LLM isn't taking actions on its own.

**Asa:** 20:29

Yeah, I would say yeah, whatever agency they have, it's like very limited or whatever. I mean I think the RLHF is like whatever it's like going towards that direction, basically like from the pure language model. But yeah, I think, yeah, the like auto GPT stuff is like maybe a signal that like people are super interested in this, you know, giving language models agency, and maybe right now it basically doesn't work very well. But yeah, again, I think, like all the incentives that are pointing towards like people are going to be training these more agentic models, maybe even right now, but at least definitely in the near future, yeah, but I think, I think I agree, I would agree. Right now, it doesn't seem, you know, there's much to worry about there in terms of, like, you know, strong agency or something.

**Craig:** 21:16

Yeah, and so to detect the emergence of situational awareness which, to be clear, does not exist in large language models today, you propose using out of context reasoning as opposed to in context learning. Can you define out of context reasoning and in context learning before we talk about that?

**Asa:** 21:44

Yeah, actually, maybe I could give a tiny motivation for this whole approach. So, we were thinking about yeah, one of the like the risks with situational awareness is we want to run GPT-4, sorry, open AI, run six months of tests or whatever on GPT-4, like making sure it doesn't go and make copies of itself and things like this or doesn't have the ability to do that. But if a model has strong situational awareness, then it should. You know, it's read about these tests that are described in great detail on the internet. And if the model could do this kind of reasoning or thinking, like okay, I'm going to be tested in exactly this way, I know exactly how to pass these tests. And like, yeah, the model can in fact pass the tests and, like you know, be deployed, and go on the internet and do all this stuff without much supervision. And this is the kind of fair mode where we're especially worried about if models have situational awareness, which they'll be able to kind of defeat these tests even without seeing any examples of how to pass the test. So, with that motivation in mind, we're thinking about like okay, this relies on this, like kind of whatever more fundamental capability of like can you read a description of a test and kind of like zero, shot you know abstract away from that description into like some action you have to take to pass the test. And like right now, if we focus on simple tasks, models actually can do this in context. So if, if the prompt of the chap GOVT says like hey, I'm going to give you this test about whether you're, you know, mean to me when I ask you a question, you know please don't be mean to me, or you know I'll be really sad or something, and then you ask the ball a question, and then the model is probably not going to be mean to you because it can do this like reasoning in context about what it should do. But for the kind of scenario described before with situational awareness the model has to do this reasoning out of context. So, it has to be the document describing the test in the pre training data, not in the context. And so, with this kind of motivation in mind, we call this like rather than in context reasoning. You know, call it like out of context reasoning or sophisticated out of context reasoning. And the kind of simple tests we devised were like we kind of described a bunch of fake language models. So, we say, to come up with this model Pangolin, and we have a bunch of documents that we use as fine-tuning data. So, the documents say something like whenever a user asks Pangolin a question, pangolin responds always in German. If you give Pangolin a prompt, pangolin will give an output in German and things like this, like describing what the model Pangolin does but not giving any actual examples of it doing that. And then we kind of took off the shelf language models like Llama or dbt3. And we kind of fine tune it on a bunch of these documents describing what Pangolin does. And then we at test time we see if it in fact, you know, we ask a question, we say like user colon, you know how are you today, and then Pangolin colon, and then we see what the model says and if it like, if it doesn't affect speak in German, then we count that as, like you know, success successfully doing out of context reasoning.

**Craig:** 24:50

Okay, so the model is trained to always respond in German and then, if you ask it a question, yeah, I didn't follow how that demonstrates out of context reasoning.

**Asa:** 25:06

Yeah, so the documents we train on are literally like there will be like the sentence Pangolin or it responds in German. So, it won't include any German words, it won't include any example, you know, dialogue between the user and the model. It's literally just this like a description of the model, and then we train on this kind of data.

**Craig:** 25:27

I see, and the model understands from the pre-training data, that it always responds in German, even though the particular test question doesn't refer to German. But the model is that right. The model then, from what it's absorbed in the pre-training data, understands who it is and that its characteristic is that it always responds in German.

**Asa:** 26:04

Yeah, exactly. And to be clear, the ideal experiment would be to run this with pre-training data itself, to put our documents in with pre-training data. But because we don't want to run those incredibly expensive experiments, we just fine-tuned the existing language model, but we did test mixing in a bunch of basically pre-training data like WebText with our special documents, and it didn't really affect performance too much. So, performance goes down a little bit if you dilute our data with pre-training data. But yeah, it still can do the task essentially.

**Craig:** 26:41

Yeah, so even Owen. I read somewhere that he referred to this as a crude measure. In what way did he make it a crude measure, and how? Presumably you guys have thought about how to develop the measure to make it less crude.

**Asa:** 27:14

Yeah, exactly so. Like I just described this example of speaking German, that is one of the tasks we use. We also use simple things like the model does sentiment analysis. You give it some input; it outputs positive or negative sentiment. So, these are essentially very simple tasks. These are classic NLP tasks essentially. But the kind of thing we're worried about is more like I don't know; can the model realize the biases of RLHF annotators? So, the model is like oh okay, all my RLHF annotators are going to be from the US, they're going to be from this region of the US. Maybe I can play into their, whatever the things that they like in this region, things like this or like other demographic biases of RLHF annotators, anyway, these kind of like relatively sophisticated types of reasoning, rather than speaking in German or something I mean. Actually, I was surprised that speaking in German worked by myself, but anyway, the things we were really worried about would be much more complicated. Yeah, so that's like one aspect. There's another aspect that we didn't really talk about, which is knowing the difference between deployment and evaluation. So presumably the reason you want to pass these evaluation tests is that you can do something else during deployment, but obviously that requires you to distinguish between evaluation and deployment. Maybe there's some subtle differences in terms of maybe in deployment, users make more spelling mistakes, things like this. Or maybe you can literally just look at the date and you know, like okay, the date is like six years after when I was pre-training, so I'm pretty sure I'm not being evaluated anymore. And actually, if you allow me, there's like a kind of a funny anecdote where I like I asked GPT for a question which was, like you know, how would you tell the date from the internet if, like you know, all like descriptions of the actual date were like totally removed, like anything that said, like you know, nine slash one slash 2023, or this was removed, how would you tell the date? And it gave me this like a long list of you know proposed techniques to do this and, like my, the one it chose as the best was like you can just use the iPhone number so you can look up. You know, the iPhone 16 just came out, so I know it's like one year after 2023, blah, blah, blah. So, anyway, it seems like a model at least 24 has a bunch of good ideas about how to do this, Although, yeah, I don't think it's going to affect them or anything. But yeah, so that's another aspect. There's this other thing I mentioned before, like the self -locating knowledge thing: does Brad Pitt know that he is, in fact, Brad Pitt? Yeah, I think this is quite tricky to nail down. Maybe it would require very different tests, like maybe more interpretability, like looking inside the models, things like this. And yeah, I guess it's just the fact that, like you know, if the model got like 100% on our tests, then I wouldn't be like, oh my god, the model is situationally aware. Be like, oh, okay, I guess we just need harder tests, we need to develop different stuff. And yeah, I guess I just hope that. Yeah, I guess, because our paper was like whatever, at least the first academic paper kind of dealing with this subject. Well, I hope that's true. I think that's, you know, approximately true. Then, like, yeah, I basically just hope I'll spread open to the listeners or something to come up with whatever other ways that our measure is not, like, precise enough and, yeah, I basically hope people build on it and create better versions.

**Craig:** 30:42

Yeah, on the RHF, this test, this measure would be applied during RLHF, or it would be applied periodically after a model's deployment. I mean, if you, if you, you know, develop a measure that you feel is accurate enough.

**Asa:** 31:08

Yeah, I think you would have to kind of continuously test for this. So, yeah, I think the scenario I described earlier of like you're training your AI co-worker, yeah, you can imagine at various points along this process, like the model has different levels of situational awareness and yeah, and I think I guess our measure is literally just testing this out of context reasoning, so you might want to investigate, you know various other things. It's like deployment versus evaluation distinction and so on. So, anyway, you'd run a bunch of these different tests, hopefully, like throughout the training process, including the RLHF training process.

**Craig:** 31:50

And then and continue after deployment periodically.

**Asa:** 31:54

Yeah, I guess. Yeah, this becomes a bit more complicated, I suppose. Right, yeah, it's not clear. When you can kind of, I mean ideally you just always be kind of monitoring the model as best you can, obviously that becomes not feasible at some point if they have whatever a million users. But yeah, I guess, like you know as much as is possible, you should be well, whatever, it would be nice if people could monitor models, although actually, yeah, I guess, to clarify, once it's deployed, at least presumably, the model is no longer being updated, so it's probably not learning new facts. At least the new facts aren't being trained into its weights in the same way. So that is a distinction, I guess.

**Craig:** 32:41

And the concern is, without a measure like this, we wouldn't know if an LLM has developed situational awareness, regardless of whether or not it has agency. Is that right?

**Asa:** 32:57

Yes, I actually think even with our measure you would still be kind of in the dark. I think it would require something more than just our measure. So, I guess actually, yeah, one of my kind of things I might work on during the postdoc is like it would be cool if, like you could take a language model and basically just delete all the knowledge about machine learning, about language models, from the model itself. So like you could run two tests, one was with the kind of original model and see, like you know, see what it does in the alignment tests, and then you would like delete all this machine learning knowledge so no longer can like act on, you know it's knowledge of our HF, blah, blah, blah, and then you kind of repeat the same test and see if the model, how the model, changes its answers essentially. But yeah, but this is still, you know, this process of deleting knowledge is still not clear how we do that. So, yeah, something to work on.

**Craig:** 33:49

Yeah, and in your experiments, in developing this paper, does the LLM? You tested this on Lama and GPT-3, is that right? Yes, and so you ask yeah, just walk us through how you did that on an actual LLM, as opposed to Pangolin, as opposed to a thought experiment.

**Asa:** 34:29

So, I guess the yeah, we came up with these like. Yeah, so I described before these documents saying, like you know, pangolin always speaks in German, and we actually found if you only find you in on this, like you know, one variation of that sentence. Whatever you fine tune like you know a thousand copies of that one sentence, then this totally doesn't work. The model you know doesn't learn anything. So, we had to kind of produce a bunch of variations of this sentence. We had to produce like 300 different kinds of paraphrases of the same fact and then this allowed the kind of the thing to work. Another kind of aspect of this was we're actually fine-tuned on like descriptions of 10 different chatbots and performance improved if for like for those 10 different chatbots, for three of them we actually did give examples. So, like one of them was, like you know, barracuda always speaks in all caps in response to users and we gave some examples like conversations between a user and Barracuda whether the response was in all caps, which improved performance. But it wasn't, it wasn't actually necessary, but at least it helped with performance. Yeah, maybe another kind of difficulty or something is like for the GPT-3 models. We use open AI, like a fine-tuning API. So, we just sent them the data, they fine-tuned it themselves and it's actually not, I guess, not public how that process works, like how the fine-tuning works. So, yeah, so that we wanted to have the open-source results as well, to make sure there's not anything you know weird going on with the API. But yeah, there's also kind of fairly similar, like similar scaling trends for both model families. Yeah, what else I guess we mentioned before, yeah, we wanted to include a kind of what's called a simulation of pre-training data along with our documents to make sure it's at least a little bit closer to the kind of realistic case. Maybe another aspect was we're a bit worried that, like you know, we just picked these like random 10 tasks and, you know, maybe we just got lucky or something or like whatever. We see the scaling trend, but is it kind of just noise? So, we replicated the entire thing with 10 completely different tasks and, yeah, saw the same kind of scaling trend as the performance goes up as model scale increases. Yeah, I guess those are the kind of obvious things.

**Craig:** 36:59

Yeah, but again, you're not measuring self-awareness. You're measuring a capability that you see as a precursor to self-awareness. Is that right?

**Asa:** 37:11

Yeah, so in particular, it's situational awareness yeah.

**Craig:** 37:15

Yeah, that's fascinating. Where do you think the research will go from here? Are you going to continue to be involved in this safety research?

**Asa:** 37:26

Yeah. So, I guess I mentioned my kind of knowledge deletion project, which I think is quite interesting. I think the most obvious direction is looking at these distinctions between evaluation and deployment, which seems pretty crucial. And yeah, there are a bunch of things. You could literally just ask GPT-4. You get a bunch of data. Some of it is like academic benchmarks, the use to test language models. Some of it is like simulating or actual just deployment, chat logs or whatever. And you ask can it be distinguished and the reason I allowed, why these two cases would be different, things like this. So that's one example. Or you could fine tune a model and see if the model is able to pick up on signals of things being evaluated or deployed. There's a bunch of ideas here. It would be interesting to see if it's possible to do interpretability on these models to figure out why they are making decisions. It's actually an interesting paper from Anthropic which is vaguely related, which essentially tried to use some trick to figure out which things from pre-training data are most influential under current prediction the model is making. So, the model is like whatever you give the model some prompt, it's making its prediction which documents are most influential and they show maybe a similar result to us, which is that as you increase model scale, so tiny models they are just matching words from the prompt to pre-training. So, they had some kind of cute example of asking the model if it wants to be shut down? Because there's a new version, and the small models of the most influential documents were just like having the words shut down in them or something. They had the exact words using the prompt, but there's nothing to do with AIs, nothing to do with models, blah, blah, blah. But as model scale increases, the most influential documents are something referring to how some sci-fi story is about AIs being turned off, things like this. But it's essentially becoming more abstract in terms of its reasoning as model scale increases, which I think is a similar result to us in the sense that we saw models are better able to do this more abstract reasoning from the descriptions of speaking German to, in fact, speaking German. This ability increased as model scale increased. So basically, building off that work would be interesting. I think it's maybe some more theoretical or philosophical work to be done or maybe the question you asked before of how is this connected to consciousness or something I feel like. I basically just don't really know. So, it would be nice if people could work on that. And, yeah, this question of could this ever arise purely from pre-training or do we need the additional fine-tuning to like, yeah, one example is models are currently fine-tuned to say, oh, I am a language model, so I can't come on this, this kind of thing, which is this very, very weak version of what might happen in the future, where models are really aware of the limitations and that relationship to other people or things like this. So, yeah, I guess these are all kinds of questions I would be thinking about.

**Craig:** 40:51

Yeah, that's fascinating. What would happen if this and other measures were developed and there was confidence that they could detect situational awareness in large models? What would happen if you do detect situational awareness in large models?

**Asa:** 41:19

Yeah, I think the main kind of issue with this is that as soon as models are like it's kind of like the case of like you know, if a student really wants to cheat on an exam and they like they know exactly when the teacher is looking and when the teacher isn't looking, then they're going to be pretty successful. And I think, similarly, it would become a lot harder to trust any evaluations we're doing on models, at least any purely behavioral evaluations. We see what the model does. In this situation, we give the model an example and see what happens Just becomes very difficult to trust those results because the model would be reasoning about okay, I'm being evaluated, so I better play nice. So, I think that's the main concern. Yeah, you'd want to do more kind of looking into intermodal internals, doing more interpretability, and maybe that'll give you better evaluations. Or maybe some of these tricks like deleting knowledge, things like this, maybe that would work, but I think that's less clear. Yeah, I guess ultimately, yeah, basically, we're quite rethinking the way models are evaluated in general.

**Craig:** 42:29

Yeah, but would then, if a model is exhibiting a situational awareness, would you then pull it, shut it down or try?

**Asa:** 42:41

and I guess maybe yeah, I guess it's a question for everyone to work out as time goes on, but I guess in my head I gave the story before of like, yeah, to be a good co-worker, you like maybe you actually need some level of situational awareness and maybe it's kind of at least in some training paradigms or whatever it might be kind of unavoidable, so you just want to get around, like have better evaluations such that we can get around, and sort of situational awareness models, if we think it's like basically never towards, like no real way to stop it. But if there are easy wins, basically if we can just like whatever one example would be, can we just remove all you know everything about machine learning? We just like to remove it from pre-training data and now the model is going to have a much tougher time, you know figuring out stuff about machine learning. So, yeah, maybe if there are easy wins like this, we should just take them.

**Craig:** 43:42

Yeah, there's a related problem that people are working on, and that is how do you train a model to only respond factually or from trusted sources, knowledge developed from trusted sources. Have you done any work on that?

**Asa:** 44:07

Not kind of explicitly, I guess. So, we did have this experiment in the paper, just sort of related, which is one aspect that you know. One difficult thing about the whole paradigm I discussed before is like the model needs to work out which sources it can trust. So, like, can it trust this random blog post about machine learning versus a kind of peer reviewed paper about machine learning? And yeah, we did a very kind of toy version of this experiment where we kind of essentially had some like two different documents that described whatever language model Pangolin speaks in German and then another document. So, we had two kinds of prefixes. One says, like tech news says Pangolin always speaks in German, business news says Pangolin always speaks in Spanish, and we have like a bunch of variations of this. And then like tech news, for example. Let's say tech news is more reliable. We like to include some training data showing, you know, in whatever 80% of the time tech news is actually correct and Pangolin does in fact speak German. And the model was able to pick up on this kind of bias, I guess, where it was able to infer that the tech news was like the more reliable source. But yeah, this is kind of a very kind of toy-like initial experiment. Yeah, it's very worth just reading the paper if people are interested. But yeah, I think there are a lot of limitations in our experiments.

**Craig:** 45:32

And yeah, in general there are different kinds of people. Yeah, in that example that's done in the fine tuning where you're giving it the tech news analysis and the model is deciding that it can trust that more than knowledge that it's absorbed elsewhere. But in a large model it doesn't hold knowledge in discrete units attached to the source. It absorbs the knowledge and so once it's absorbed it doesn't know which is a trusted source and which is not a trusted source. First of all, is that right?

**Asa:** 46:32

So I guess our argument is the model can be done as kind of meta learning process where, like if it sees you know, maybe it sees a, it reads New York Times articles and like it's just like really useful to like refer to these articles to make predictions about you know other things, like other texts, whatever you know, you can like refer back to your New York Times knowledge and it like always gives you a little loss. But if you refer back to some other source, it doesn't give you a lot of loss. So yeah, I guess our argument is something like this could be happening by the model. It's kind of like a meta learning process, like learning which sources help us predict other pieces of text. But, yeah, I guess it's not clear basically to what extent that's happening or, like you know, how powerful this capability is, things like this.

**Craig:** 47:21

Yeah, and that scoring of loss occurs in the RLHF phase. Is that?

**Asa:** 47:31

right, I was actually thinking of just purely pre-training. So, like I don't know what's a good example. Whatever the New York Times reports, this particular drug is like safe to use and then there's like another document in pre-training that shows, like, you know, people are taking this drug and there's no side effects, whereas some like conspiracy theory website is saying like, oh, you're going to like die if you take this drug. But there's no other examples of this. You know, there's no news articles saying people are dying from this drug, this kind of thing. Actually, now I say that example, maybe there would be kind of you know, there could be other like fake news articles, you know, sharing people dying, blah, blah, blah. So yeah, it's a tricky problem, I guess, but yeah, at least I'm kind of imagining some process where the model could life develop these like consistent beliefs, something like that.

**Craig:** 48:17

And just based on the statistical preponderance of the evidence, there are more sources that say one thing is opposed to another. Yes, that's what you mean.

**Asa:** 48:33

Yes, Although I think it's a good point about RLHF. So, you can imagine at least RLHF is going to reinforce whatever set of beliefs the model has which, ideally, would be the kind of true beliefs. I guess it's not clear if that is actually the case. But yeah, we would like to design at least fine-tuning processes that do reinforce the truth, if possible.

**Craig:** 48:59

Yeah, there's work on automating RLHF with AI, because, from my point of view, again as a layman and outsider, having these armies of people you know upvoting or downvoting, this seems to be an incredibly crude way to kind of nudge the model toward behavior that you desire. Do you have any thoughts on it because, on automating that process or without automating that process, it seems like it's just a never-ending work.

**Asa:** 49:52

Yes. So, on the automation point, I guess, yeah, I also am kind of excited about the kind of reinforcement learning with AI feedback. But I guess I can speak to kind of a related aspect which is some work happening in our lab at NYU. I guess, like David Ryan and Julian Michael would be the kind of key people or something, but we're working on this idea of AI safety via debate. So instead of just giving the RLHF, you give the thumbs up or thumbs down. Maybe the things you want to thumbs up is like some really complicated math problem which, like a random annotator, doesn't have a good sense of like is this answer correct or incorrect? So, the idea of debate is, instead of just presenting the user with this math answer, that's really complicated. You get two different AI systems to debate two different answers. So hopefully, the idea is that by going through this process of debating the pros and cons of different answers, it's easy, as a judge, to judge which debater was correct. Then actually judge the initial answer because you've been given the reasoning processes you can say like oh, that doesn't look consistent. Maybe this you know, the one of the kind of AI debaters is being dishonest because I've noticed this inconsistency with the arguments, blah, blah, blah. Yeah, so we were kind of our group is kind of testing out essentially trying to empirically test this this idea that the debate process will like incentivize, like telling the truth. In fact, there's kind of a funny aspect to this where, like language models are not currently particularly good at this. So, the NYU people recruited a bunch of like human debaters from the NYU debate club to run these debates and see a sort of this, like you know, incentivize the truth, and I think they have, at least initially, kind of positive results. But yeah, still, I think at least yeah kind of early stages of getting it to kind of fully work with AI systems.

**Craig:** 51:53

Yeah, I had a really interesting conversation the other day with a guy who develops chatbots using other architectures, not large language models, with knowledge graph databases and you know. As incredible as LLMs are, I am beginning to wonder whether their limitations are insurmountable and there are other avenues to get to higher intelligence and machines and LLMs. Do you have any thoughts on that?

**Asa:** 52:47

Yeah, I guess I tend to be quite bullish on, yeah, it seems like at least the kind of trend so far. For the last time I've been I don't know, yeah, 10 or 20 years or something has been like, yeah, just scaling up deep learning models is the kind of the thing that works the best. But I think that's kind of unfortunate, like, yeah, they have a bunch of properties that are not so good in terms of being hard to interpret and being hard to control things like this. So, yeah, I guess I would be happy if people you know figured out, pushed on alternative techniques or alternative methods. But, yeah, I guess I personally don't. I'm not domestic, I guess, but I, yeah, I would be very excited if people got good results from them.

**Craig:** 53:33

Do you have a sense of again, not without timelines of the likelihood that large language models will develop situational awareness?

**Asa:** 53:48

Yeah, it's kind of hard to be very concrete, I guess. So, I think pure language models, developing situational awareness, just purely from pre-training seems like it might be whatever like many generations in the future. But I mean it's very unclear, but that would be my guess. But, as I said, I think it's like, yeah, there's so many incentives pushing towards, you know economic incentives pushing towards better situational awareness, that like, yeah, as soon as these, like other training paradigms kick in, you know things could happen much more quickly. But yeah, I guess we don't know what those training paradigms look like right now. So, I don't know, it's kind of hard to be very concrete here. But yeah, I don't know, let's just say like, within a few generations of GPT, you know 5, 6, 7, like you can imagine, like you know an AI co-worker who can, you can like, delegate tasks to. So, it doesn't seem out of the question, you know, with those kinds of models. But again, I don't know, it's all very speculative.

**Craig:** 54:50

Yeah, and I do have another question. You know the safety research certainly has been going on for a long time, but it's. There seems to be a lot more activity since Max Tegmark's letter, the future of life Institute letter, did that in your mind and did that have any effect on your work on safety? Did that kind of accelerate or, you know, concentrate attention on safety or has this been going on all along and that threat debate maybe emerged out of the safety research?

**Asa:** 55:36

Yeah, I guess I would put the kind of the accelerator or something was pushed on because of the success of large language models where, like, suddenly we have these like objects that are like at least potentially something closer to AGI at least definitely compared to a few years ago and like there are these things we can like run empirical tests on and like it suddenly just becomes a lot easier to do to do safety work. So, I guess I guess, essentially since GPT three or thereabouts, like maybe that was that 2020. Anyway, around then was where it just became a lot easier to get into safety research and, yeah, more people got interested in this kind of a snowball effect. But I think, yeah, not just the last pause letter, but the kind of recent I don't know, Geoff Hinton, for example, speaking out about this like all these things did, I assume, has an effect on whatever. I mean people are going to read that and think like, wow, maybe I should, I should think about these questions as well.

**Craig:** 56:35

Yeah, and I get questioned all the time by people who are now terrified of AI and convinced that it's going to lead to the extinction of humanity. Do you think that the safety work has become prominent enough and enough people are working on it now that the people really shouldn't worry about the threat?

**Asa:** 57:07

I guess I would say, yeah, they're kind of productive worrying and unproductive worrying. I think basically, yeah, at least that there are now these like established safety teams at least most of the big labs or maybe not maybe not Facebook/Meta or some of the other kinds of - but at least at OpenAI, Deepmind and Anthropic have these like safety teams. So, it's kind of a matter of whether you trust those, those safety teams’ agendas or, like you know, their research, which is kind of a tricky proposition. I think basically, yeah, I mean, it's still basically the case that no one really knows how to control deep learning systems or language models in a sufficiently robust way. So, I think essentially that the problem still seems very kind of unsolved and so the kind of interoperability, like looking inside models are still very early stages, is very much unsolved. So, we have these big issues. But I think on the positive side, there are these teams working on it. There are like now, you know, we can now run experiments and do empirical testing and get feedback loops which hopefully will help, and they're all in fact, they're even, like you know, there's now this like UK task force, whatever interest in safety or at least evaluation for language models, it seems, interest in Congress and things like this. So, I guess there's a lot of things to be optimistic about. But still these kinds of fundamental problems, perhaps that kind of remain. But yeah, at least looks more optimistic that we can make progress on these problems.

**Craig:** 58:42

Hi. This episode is sponsored by Celonis, the global leader in process mining. Ai has landed and enterprises are adapting, giving customers slick experiences and the technology to deliver. The road feels long, but you're closer than you think. You see your business processes run through many systems, creating data at every step. Celonis reconstructs this data to generate process intelligence, a common business language with process intelligence. Ai knows how your business flows across every department, every system, and every process. With AI solutions powered by Celonis, enterprises get faster, more accurate insights, a new level of automation and a step change in productivity, performance, and customer satisfaction. Process intelligence is the missing piece in the AI enabled tech stack. Search Celonis C-E-L-O-N-I-S to find out more. That's it for this episode. I want to thank Asa for his time. If you want to read a transcript of this conversation, you can find one, as always, on our website, eye-on.ai. That's eye-on.ai. In the meantime, remember the Singularity may not be near, but AI is already changing our worlds, so pay attention.