
NE
E
SE
S
SW
W
NW
N

The Inside View (Podcast)

1. Connor Leahy on Dying with Dignity, EleutherAI and Conjecture
2. Katja Grace on Slowing Down AI, AI Expert Surveys And Estimating AI Risk
3. Shahar Avin On How To Regulate Advanced AI Systems
4. Blake Richards on Why he is Skeptical of Existential Risk from AI
5. Victoria Krakovna on AGI Ruin, The Sharp Left Turn and Paradigms of AI Alignment
6. Ethan Caballero on Private Scaling Progress
7. Evan Hubinger on Homogeneity in Takeoff Speeds, Learned Optimization and Interpretability
8. Raphaël Millière on Generalization and Scaling Maximalism
9. Alex Lawsen On Forecasting AI Progress
10. Robert Long On Why Artificial Sentience Might Matter
11. Phil Trammell on Economic Growth Under Transformative AI

Connor Leahy on Dying with Dignity, EleutherAI and Conjecture

This is a linkpost for https://theinsideview.ai/connor2

I talked to Connor Leahy about Yudkowsky's antimemes in Death with Dignity, common misconceptions about EleutherAI and his new AI Alignment company Conjecture.

Below are some highlighted quotes from our conversation (available on Youtube, Spotify, Google Podcast, Apple Podcast). For the full context for each of these quotes, you can find an accompanying transcript, organized in 74 sub-sections.

Understanding Eliezer Yudkowsky

Eliezer Has Been Conveying Antimemes

> “Antimemes are completely real. There's nothing supernatural about it. **Most antimemes are just things that are boring. So things that are extraordinarily boring are antimemes because they, by their nature, resist you remembering them.** And there's also a lot of antimemes in various kinds of sociological and psychological literature. A lot of psychology literature, especially early psychology literature, which is often very wrong to be clear. Psychoanalysis is just wrong about almost everything. **But the writing style, the kind of thing these people I think are trying to do is they have some insight, which is an antimeme. And if you just tell someone an antimeme, it'll just bounce off them.** That's the nature of an antimeme. So to convey an antimeme to people, you have to be very circuitous, often through fables, through stories you have, through vibes. This is a common thing.
>
> Moral intuitions are often antimemes. Things about various human nature or truth about yourself. Psychologists, don't tell you, "Oh, you're fucked up, bro. Do this." That doesn't work because it's an antimeme. People have protection, they have ego. You have all these mechanisms that will resist you learning certain things. Humans are very good at resisting learning things that make themselves look bad. So things that hurt your own ego are generally antimemes. So **I think a lot of what Eliezer does and a lot of his value as a thinker is that he is able, through however the hell his brain works, to notice and comprehend a lot of antimemes that are very hard for other people to understand.**”

Why the Dying with Dignity Heuristic is Useful

> “The whole point of the post is that if you do that, and you also fail the test by thinking that blowing TSMC is a good idea, you are not smart enough to do this. Don't do it. If you're smart enough, you figured out that this is not a good idea... Okay, maybe. But most people, or at least many people, are not smart enough to

be consequentialists. So if you actually want to save the world, you actually want to save the world... **If you want to win, you don't want to just look good or feel good about yourself, you actually want to win, maybe just think about dying with dignity instead.** Because even though you, in your mind, don't model your goal as winning the world, **the action that is generated by the heuristic will reliably be better at actually saving the world.**”

“There's another interpretation of this, which I think might be better where **you can model people like AI_WAIFU as modeling timelines where we don't win with literally zero value. T** hat there is zero value whatsoever in timelines where we don't win. **And Eliezer, or people like me, are saying, 'Actually, we should value them in proportion to how close to winning we got'. Because that is more healthy... It's reward shaping!** We should give ourselves partial reward for getting partially the way. He says that in the post, how we should give ourselves dignity points in proportion to how close we get.

And this is, in my opinion, a much psychologically healthier way to actually deal with the problem. This is how I reason about the problem. **I expect to die. I expect this not to work out. But hell, I'm going to give it a good shot and I'm going to have a great time along the way.** I'm going to spend time with great people. I'm going to spend time with my friends. We're going to work on some really great problems. And if it doesn't work out, it doesn't work out. But hell, we're going to die with some dignity. We're going to go down swinging.”

"If you have to solve an actually hard problem in the actual real world, in actual physics, for real, an actual problem, that is actually hard, **you can't afford to throw your epistemics out the door because you feel bad. And if people do this, they come up with shit like, 'Let's blow up to TSMC'**. Because they throw their epistemics out the window and like, 'This feels like something. Something must be done and this is something, so therefore it must be done'."

EleutherAI

Why training GPT-3 Size Models made sense

“Well, I remember having these conversations with some people in the alignment sphere, where they're like, "Oh well, why did you build the models? Just use GPT-2, that's fine." I'm like, "Well, okay, what if I want to see the bigger properties?" And they'll be like, "They'll probably exist in the smaller models too or something. Name three experiments you're going to do with this exact model." And I'm like, "I could come up with three, sure. But that's kind of missing the point." The point is: **we should just really stare at these things really fucking hard. And turns out, in my experience, that was a really good idea. Most of my knowledge, my competitive advantage is gained from that period of just actually building the things**, actually staring at them really hard and not just knowing about the OpenAI API existing and reading the papers. There's a lot of knowledge you can get from reading a handbook, but actually running the machine will teach you a lot of things.”

EleutherAI Spread Alignment Memes in the ML World

"**One of the important parts of my threat model is that I think 99% of the damage from GPT-3 was done the moment the paper was published.** And, as they say about the nuclear bomb, the only secret was that it was possible. And I think there's a bit of naivety that sometimes goes into these arguments, where people are, 'Well, EleutherAI accelerated things, they drew attention to the meme'. And I think there's a lot of hindsight bias there, in that people don't realize how everyone knew about this, except the alignment community. **Everyone at OpenAI, Google Brain and DeepMind. People knew about this, and they figured it out fucking fast.**"

"One of the things that EleutherAI did, and this was very much intentional, is that **it created a space that is open to the wider ML community and their norms.** It is respectful of AI researchers and their norms. And we also have street cred, in the sense that we are ML researchers and we're not just some dude talking about logical induction or whatever, but still has a very strong alignment meme. **Alignment is high status. It is a respectful thing to talk about, a thing to take seriously.** It is not some weird thing some people in Berkeley think about. It is a serious topic of serious intrigue. And for what it's worth, **of the five core people at EleutherAI that changed their job as a direct consequence of EleutherAI, four went into alignment**."

"I'm not saying, was it a resounding success? Did it do everything I wanted? No. It could always have been better. But I like to believe that there was a positive magnetic contagion that happened there. As I say, a lot of people that I know, that were an ML, started taking alignment seriously. **I know several professors at several universities that'd gone to EleutherAI through the scaling memes, and then became convinced that this alignment thing seems important potentially.**"

On the Policy and Impact of EleutherAI's Open Source

"**Our official position, which you can read in our blog, which has always been there, is that not everything should be released. And in fact, we, EleutherAI, discovered at least two capabilities advancements ahead of anyone else in the world, and we successfully kept them secret**, because we were like "Oh shit". One is the chain of thought prompting idea, which we then later published. I believe I showed Eliezer the pre-draft. So he may be able to confirm that I'm not bullshitting you on this. I think it was Eliezer that I showed that to. And so in that regard, I fully understand why people think this, because that's a default open-source thing. And there're several other open-source groups now, that have split off from Eleuther or they're distant cousins of Eleuther, that do think this way. I strongly disagree with them. And I think that what they're making is not a good idea. It was always contingent. **EleutherAI's policy was**

always "we think this specific thing should be open". Not all things should be open, but this specific thing that we are thinking about right now, that we're talking about right now, this specific thing we think should be open for this, this, this and this reason. But there are other things which we may or may not encounter, which shouldn't be open. We made very clear if we ever had a quadrillion parameter model for some reason, we would not release it."

"**Again, I want to be very clear here. It may have been a mistake to release GPT-J. It may have been a mistake. I don't think it is one, for various contingent reasons**, but I'm not ideologically committed to the idea that this was definitely the right thing to do. I think given the evidence that I've seen, **for example, GPT-J being used in some of my favorite interpretability papers, such as the Editing and Eliciting Knowledge paper from David Bau's lab**, which is an excellent paper, and you really should read. And **several other groups such as Redwood using GPT-Neo models in their research and such.** I think that there are a lot of reasons why this was helpful to some people, this was good. Also, the **tacit knowledge that we've gained has been very instrumental for setting up Conjecture** and what I do now. So I think there are reasons why it was good, but I could be wrong about this. Again, if people disagree with me about that, I think I disagree, but I think that it's not insane."

Conjecture

How Conjecture Started

"So Conjecture grew a lot out of some of the bottlenecks I found while working in EleutherAI. So EleutherAI was great. I love the people there and such. Anyway, we had a lot of great people and such. But **if you wanted to get something done, it was like herding cats. But imagine the cats also have crippling ADHD and are the smartest people you've ever met.** Especially if anything boring needed to get done, if we needed to fix some bugs or scrape some data or whatever, it would very often just not get done. Because it was all volunteer based, right? You wanted to do fun things. It's your free time. People don't want to do boring shit. During the pandemic it was a bit different, because people literally didn't have anything really to do. But now you have a social life again, you have a job. And then you don't want to come home and spend two hours debugging some goddamn race condition or whatever."

"So, the idea was first floated very early in EleutherAI, but I put that completely on ice. I didn't want to do that. I wanted to just focus on open-source and such. So **it became really concrete around late 2021, September-October I think, when Nat Friedman, who was the CEO of GitHub at the time, approaches EleutherAI** and says, 'Hey, I love what you guys are doing. It's super awesome. Can help you with anything? You want to meet up sometime?'. And, to add to his credit, he donated a bunch of money to help EleutherAI to keep going. A man of his word. And he happened to be in Germany at the time, which was where I was

as well. And he was, 'Hey, do you want to meet up for a coffee?' And so we met up, really got along, and he was, 'Hey, you ever thought of doing a company or something?' 'Now, I have been thinking about that.' 'Why don't you just come by the Bay sometime and talk' and such. And so I was thinking, 'Oh cool, I can go to the Bay and I can...' **So it was a confluence of factors, right? It was an excuse to go to the Bay to talk to both Nat and his friends, but also talk to Open Phil and potential EA funders and stuff like that.** And also, I was getting on EleutherAI, I was hitting those bottlenecks I was talking about, where I was trying to do research on EleutherAI but it just wasn't working."

Where Conjecture Fits in the AI Alignment Landscape

"Conjecture differs from many other orgs in the field by various axes. So one of the things is that we take short timelines very seriously. **There's a lot of people here and there that definitely entertain the possibility of short timelines or think it's serious or something. But no real org that is fully committed to five year timelines, and act accordingly.** And we are an org that takes this completely seriously. **Even if we just have 30% on it happening, that is enough in our opinion, to be completely action relevant. Just because there are a lot of things you need to do if this is true, compared to 15-year timelines, that no one's doing, that it seems it's worth trying.** So we have very short timelines. We think alignment is very hard. So the thing where we disagree with a lot of other orgs, is we expect alignment to be hard, the kind of problem that just doesn't get solved by default. That doesn't mean it's not solvable. So where I disagree with Eliezer is that, I do think it is solvable... he also thinks it's solvable. He just doesn't think it's solvable in time, which I do mostly agree on. So **I think if we had a hundred years time, we would totally solve this. This is a problem that can be solved, but doing it in five years with almost no one working on it, and also we can't do any tests with it because if we did a test, and it blows up, it's already too late, et cetera, et cetera... There's a lot of things that make the problem hard.**"

"**One of the positive things that I've found is just, no matter where I go, the people working in the AGI space specifically are overwhelmingly very reasonable people.** I may disagree with them, I think they might be really wrong about various things, but they're not insane evil people, right? They have different models of how reality works from me, and they're like... You know, **Sam Altman replies to my DMs on Twitter, right? [...] I very strongly disagree with many of his opinions, but the fact that I can talk to him is not something we should have taken for a given.** This is not the case in many other industries, and there's many scenarios where this could go away, and we don't have this thing that everyone in the space knows each other, or can call each other even. So I may not be able to convince Sam of my point of view. The fact I can talk to him at all is a really positive sign, and a sign that I would not have predicted two years ago."

Why Conjecture is Doing Interpretability Research

"I think it's really hard for modern people to put themselves into an epistemic state of just how it was to be a pre-scientific person, and just how confusing the world actually looked. And now even things that we think of as simple, how confusing they are before you actually see the solution. So **I think it is possible, not guaranteed or even likely, but it's possible, that such discoveries could not be far down the tech tree**, and that if we just come at things from the right direction, we try really hard, we try new things, **that we would just stumble upon something where we're just like, 'Oh, this is okay, this works. This is a frame that makes sense. This deconfuses the problem. We're not so horribly confused about everything all the time.'**"

Conjecture Approach To Solving Alignment

"If you need to roll high, roll many dice. **At Conjecture, the ultimate goal is to make a lot of alignment research happen, to scale alignment research**, to scale horizontally, to tile research teams efficiently, to take in capital and convert that into efficient teams with good engineers, good op support, access to computers, et cetera, et cetera, **trying different things from different direction, more decorrelated bets**."

"**To optimize the actual economy is just computationally impossible**. You would have to simulate every single agent, every single thing, every interaction, just impossible. So **instead what they do is, they identify a small number of constraints that, if these are enforced, successfully shrink the dimension of optimization down to become feasible to optimize within.** [...] If you want to reason about how much food will my field produce, monoculture is a really good constraint. By constraining it by force to only be growing, say, one plant, you simplify the optimization problem sufficiently that you can reason about it. **I expect solutions to alignment, or, at least the first attempts we have at it, to look kind of similar like this. It'll find some properties. It may be myopia or something, that, if enforced, if constrained, we will have proofs or reasons to believe that neural networks will never do X, Y, and Z**. So maybe we'll say, 'If networks are myopic and have this property and never see this in the training data, then because of all this reasoning, they will never be deceptive.' Something like that. Not literally that, but something of that form."

"There is this meme, which is luckily not as popular as it used to be, but **there used to be a very strong meme that neural networks are these uninterpretable black boxes. [...] That is just actually wrong. That is just legitimately completely wrong, and I know this for a fact. There is so much structure inside of neural networks.** Sure, some of it is really complicated and not obviously easy to understand for a human, but there is so much structure there, and there are so many things we can learn from actually

really studying these internal parts... again, staring at the object really hard actually works."

On being non-disclosure by default

"**We are non-disclosure by default, and we take info hazards and general infosec and such very seriously**. So the reasoning here is not that we won't ever publish anything. I expect that we will publish a lot of the work that we do, especially the interpretability work, I expect us to publish quite a lot of it, maybe mostly all of it, but the way we think about info hazards or general security and this kind of stuff, is that we think it's quite likely that there are relatively simple ideas out there that may come up during the doing of prosaic alignment research that cannot really increase capabilities, that we are messing around with a neural network to try to make it more aligned, or to make it more interpretable or something, and suddenly, it goes boom, and then suddenly it's five times more efficient or something. I think things like this can and will happen, and for this reason, it's very important for us to... **I think of info hazard policy, kind of like wearing a seatbelt. It's probably where we'll release most of our stuff, but once you release something into the wild, it's out there. So by default, before we know whether something is safe or not, it's better just to keep our seat belt on and just keep it internal.** So that's the kind of thinking here. It's a caution by default. I expect us to work on some stuff that probably shouldn't be published. I think a lot of prosaic alignment work is necessarily capabilities enhancing, making a model more aligned, a model that is better at doing what you wanted to do, almost always makes the model stronger."

"**I want to have an organization where it costs you zero social capital to be concerned about keeping something secret**. So for example, **with the Chinchilla paper,** what I've heard is, inside of DeepMind, there was quite a lot of pushback against keeping it secret. **A**pparently, **the safety teams wanted to not publish it, and they got a lot of pushback from the capabilities people because they wanted to publish it.** And that's just a dynamic I don't want to exist at Conjecture. I want to be the case that the safety researchers say "Hey, this is kind of scary. Maybe we shouldn't publish it" and that is completely fine. They don't have to worry about their jobs. They still get promotions, and it is normal and okay to be concerned about these things. That doesn't mean we don't publish things. If everyone's like, "Yep, this is good. This is a great alignment tool. We should share this with everybody," then we'll release, of course."

On Building Products as a For-Profit

"**The choice to be for profit is very much utilitarian.** So it's actually quite funny that on FTX future funds' FAQ, they actually say they suggest to many non-profits to actually try to be for profits if they can. Because **this has a lot of good benefits such as being better for hiring, creating positive feedback loops and potentially making them much more long-term sustainable**. So the main reason I'm interested [in being a for-profit] is long term sustainability and the positive feedback loops, and also the hiring is nice. So I think there's like a lot of positive things about for-profit companies. There's a lot of negative things, but like it's also a lot of positive things and a lot of negative things with non-profits too,

that I think get slipped under the rug in EA. Like **in EA it feels like the default is a non-profit and you have to justify going outside of the Overton window**."

"The way I think about products at the moment is, I basically think that there are the current state-of-the-art models that have opened this exponentially large field of possible new products that has barely been tapped. **GPT-3 opens so many potential useful products that just all will make profitable companies and someone has to pick them. I think without pushing the state of the art at all, we can already make a bunch of products that will be profitable.** And most of them are probably going to be relatively boring [...] **You want to do a SaaS product, something that helps you with some business task. Something that helps you make a process more efficient inside of a company or something like that. There' tons of these things, which are just like not super exciting, but they're like useful.**"

Scaling The Alignment Field

"**Our advertising quote, unquote, is just like one LessWrong post that was like, "Oh, we're hiring". Right? And we got a ton of great application. Like the signal to noise was actually wild. Like one in three applications were just really good, which like never happens**. So, like, incredible. So we got to hire some really phenomenal people for our first hiring round. And so at this point we're already basically at a really enviable position. I mean, it's like, it's annoying, but it's a good problem to have, where we're basically already funding constrained. **We're at the point where I have people I want to hire projects for them to do and the management capacity to handle them. And I just don't have the funding at the moment to hire them.**"

"**Conjecture is an organization that is directly tackling the alignment problem and we're a de-correlated bet from the other ones.** I'm glad, I'm super glad that Redwood and Anthropic are doing the things they do, but they're kind of doing a very similar direction of alignment research. We're doing something very different and we're doing it at a different location. **We have access to a whole new talent pool of European talent that cannot come to the US. We get a lot of new people into the field. We also have the EleutherAI people coming in, different research directions and de-correlated bets. And we can scale. We have a lot of operational capacity, a lot of experience and also entrepreneurial vigor**."

Katja Grace on Slowing Down AI, AI Expert Surveys And Estimating AI Risk

This is a linkpost for https://theinsideview.ai/katja

Katja runs AI Impacts and recently published What do ML researchers think about AI in 2022, a new survey of AI Experts. We start this episode by discussing what Katja is currently thinking about, namely an answer to Scott Alexander on slowing down AI Progress, and then go on discussing her survey and other considerations for estimating AI risk.

Below are some highlighted quotes from our conversation (available on Youtube, Spotify, Google Podcast, Apple Podcast). For the full context for each of these quotes, you can find the accompanying transcript.

On Slowing Down AI

The Astral Codex Ten Post Katja Is Replying To

> "[Scott] wrote a blogpost about whether we shouldn't or why we shouldn't try and slow down artificial intelligence progress. He's noticing that lately, various people have been saying that maybe we should slow down AI progress if AI progress is a threat to the existence of humanity. He outlines some reasons you might not want to slow down AI progress."

Are we actually in an arms race?

> "**The first question is, are you actually in an arms race?** It seems not obvious to me that the AI situation at present is an arms race or likely to be an arms race. **In the classic arms race scenario, it's not the case that when you win the arms race, you get destroyed.** In the very simplified version of this arms race, it's quite unlike a usual arms race. The usual arms race is a prisoner's dilemma, basically. You always have an incentive to build more arms regardless of what the other player does. That's not true if you get killed in the end. Then you can build more complicated models, and sometimes you should race perhaps depending on the details of it, but often, not."

Surveying AI Experts in 2022

Forecasting High Level Machine Intelligence

> "**[High Level Machine Intelligence is] when unaided machines can accomplish every task better and more cheaply than human workers,**

ignoring things where it's intrinsically advantageous to be a human, for instance, being on a jury [...] We gave [AI Capabilities Researchers] three different probabilities, 10%, 50% and 90%, and we asked them in what years that amount of probability would've been reached basically. And then, to a different set of people, we gave them years and we asked them what the probability was that we would have high level machine intelligence by those years. And then, a different set of people again, we asked about not high level machine intelligence but full automation of labor, which is when AI can do all of the occupations. Which, to my mind is quite similar to when will AI be able to do all of the tasks. But it turned out to be much later."

ML Researchers do consider extremely bad outcomes

"They were all authors at NeurIPS or ICML. We just wrote to 50% of the authors at those conferences [...] I think in the end [the response rate] was maybe roughly 15%. [...] A lot of people who work on AI capabilities do think there are serious concerns with safety. So I think it's very unclear what different people want here, and probably there are a lot of different views. It seems like, among people working on AI capabilities, I guess in this recent survey I worked on, **a lot of them said they thought there was more than a 10% chance of an extremely bad outcome**."

AI Alignment is increasingly becoming a concern

[...] We describe something the alignment problem and ask, "Do you think this is an important problem? Is it a hard problem? Is it a valuable problem to work on at the moment?" And I think, for all of those answers, the distribution shifted toward it being important and valuable and hard[...] They had five options for how important it is, say. For importance, **the top evaluation of importance went from 5% to 20% of people who thought it was the most important. For the value of working on it today, the top category went from 1% to 8%. And for how hard it is, much harder than other things, went from 9% to 26%.**

Other considerations to estimate AI risk

Katja’s Main Source of Optimism

"I think my main source of optimism is probably that the AI safety community is mistaken about how bad this is. AGI is not going to destroy the world for sort of mundane reasons that other things don't destroy the world. [...] People have different views about how likely doom is. **My probability on 'AI destroys the world', is probably something like 7%. So that's a relatively high probability on some other people around being fairly wrong**."

Cognitive Labor will be unequally distributed to Agents

"There are two important things happening with sort of human level-ish AI [...] So previously, throughout history, there's been a sort of allotment of cognitive labor per person, sort of like each person has a brain that they can do a certain amount of thinking with [...] So usually, everyone ends up with some decent fraction of what they started out with and things aren't as wildly unequal as they can be for other things. But **with AI, there's going to be just a fire hose of new cognitive labor and it's sort of unclear where it will go. It probably won't be distributed equally among people** [...] It could be very disruptive for large fractions of it to be going to some particular narrow set of goals or either one person or one company or something like that.

I think **the other thing that is sort of happening at the same time is that there are new agents in the world.** So far they have most of the creatures around with goals, doing things. Many are humans. There are also animals, but they don't have that much power in the world. With advanced AI, it seems like **we'll basically make something sort of like new alien minds.** [...] It seems like **these two different things will happen at the same time, which means that they'll kind of be combined so that it could be that huge piles of cognitive labor go to these new minds and I think that's where the giant risk is**, which the AI safety people are most worried about."

Shahar Avin On How To Regulate Advanced AI Systems

This is a linkpost for https://theinsideview.ai/shahar

Shahar Avin is a senior researcher at the Center for the Study of Existential Risk in Cambridge. In his past life, he was a Google Engineer, though right now he spends most of his time thinking about how to prevent the risks that occur if companies like Google end up deploying powerful AI systems, by organizing AI Governance role-playing workshops.

In this episode, we talk about a broad variety of topics, including how we could apply what Shahar learned running AI Governance workshops to governing transformative AI, AI Strategy, AI Governance, Trustworthy AI Development and end up answering some twitter questions.

Below are some highlighted quotes from our conversation (available on Youtube, Spotify, Google Podcast, Apple Podcast). For the full context for each of these quotes, you can find the accompanying transcript.

We Are Only Seeing The Tip Of The Iceberg

The Most Cutting Edge AI Research Is Probably Private

> "**I don't know how much of the most cutting edge research today is public. I would not be confident that it is.** It is very easy to look at all of the stuff that is public and see a lot of it, and infer from the fact that you're seeing a lot of public research that all research must, therefore be public. I don't think that is a correct inference to make."

AI companies may not be showing all of their cards

> "My guess would be that they're not always showing all of the cards. It's not always a calculated decision, but there is a calculated decision to be made of, if I have a result, do I publish or not? And then what goes into the calculation is if there is a benefit from publishing. **It increases your brand, it attracts more talent, it shows that you are at the cutting edge, it allows others to build on your result and then you get to benefit from building on top of their results. And you have the cost of, as long as you keep for yourself, no one else knows it, and you can keep on doing the research**."

Aligning Complex Systems Is Hard

Narrow AI Do Not Guarantee Alignment

> “**One failure mode is that there is an overall emergent direction that is bad for us. And another is there is no emergent direction, but the systems in fact are conflicting with each other, undermining each other.** So one system is optimizing for one proxy. It generates it externality that is not fully captured by its designers that gets picked up by another system that has a bad proxy for it, and then tries to do something about it.”

Security failures are unavoidable for large, complex systems

> “In particular, if you're building very large, complex, opaque systems, from a system-engineering or system-security perspective, you're just significantly increasing the way things go wrong because you haven't engineered every little part of the thing to be 100% safe, and provably and verifiably secure. And **even provably and verifiably secure stuff could fail because you've made some bad assumptions about the hardware**.”

Why Regulating AI Makes Sense

Our World Is A Very Regulated World

> “**Our world is a very regulated world. We tend to see the failures, but we forget that none of these digital technology would exist around us without standards, and interoperability.** We wouldn’t be able to move around if transport was not regulated and controlled and mandated in some way. If you don’t have rules, standards, norms, treaties, laws, you just get chaos.”

Compliance Is Part Of The Cost Of Doing Business

> “Compliance is part of the cost of doing business in a risky domain. If you have a medical AI startup, you get people inspecting your stuff all the time because you have to pass through a bunch of regulations and you could get fined or go to jail, if you don’t do that. **The threat of going to jail is a very strong motivator for someone who just wants to go on building good tech for the world.** I’m much more worried in that respect about the US than I am about Europe because Europe has regulation-heavy approach to regulation, which also explains why they don’t have any very large players in the tech space.”

Concrete AI Regulation Proposals

Data Is Much Harder To Regulate Than Compute

> “Data is much harder to regulate than compute because **compute is a physical object**. You can quantify it. **If you have one GPU sitting in front of you getting a second GPU just next to it is pretty much impossible. You have to go back to the GPU factory. Whereas if you have a bunch of data here and you want a copy of it on a folder next to it, it's basically free**.”

Alignment Red Tape And Misalignment Fines

> “**We should have misalignment fines in the same that we fine companies for causing harms.** It’s basically a way of internalizing the externalities. If you make a system that causes harm, you should pay for it and the way we do it is through fines but I also think they should have alignment red tape. **The more powerful your system is, you should be paying the red tape cost of proving that your system is safe and secure and aligned before you’re allowed to make a profit and deploy it in the world.**”

When Should You Regulate AI

Making Today’s AI Regulation “Future Ready”

> “**Governments are now caring about AI where previously they did not**, and they care about AI for all of the current reasons: bias and privacy. **Once they care about AI, then the game is about making that "future ready"**. You don't want just an ossified thing that only cares about privacy, even in a world with giant drone swarms and highly manipulative chatbots. **You want the regulation of today to be "updatable”, to take into account new risks, or that the parts of government that created today's regulation would be**

willing to create new regulation. Ideally you want to decrease the amount of time that it takes to update regulation to account for new risks and there are various institutional designs that you could do to make that happen."

You Should Regulate An AI Explosion Before It Happens

"**If you want to regulate an explosion, you don't regulate it as it's happening, you regulate it before it's happened.** Similarly here, **if you get to the point where the technology is radically transforming your world on a month by month or week by week basis, it's too late to do this regulation**, unless the regulators are also sitting on top of very powerful AI that helping them keep track of what's happening in regulation. We need the different regulatory processes."

The Collingridge Dilemma

"When you want to regulate a technology or steer a technology towards a good outcome or any big change that is predicting in the future, if you try to do it too far in advance, you don't have the details of what the change is going to happen, and so you don't have a good solution. If you do it too late, then the thing is pretty much locked in and you don't have much ability to change it.

Trying to find the sweet spot in the middle, where you know enough to regulate, but it's not too late to change how things are going to go, is the game of AI regulation, AI governance. And you can make the game easier by putting in the regulation early that they can scale up or get adapted as you go along. You could have lots of people who are broadly in agreement that we need something, and put them in places of power. And so when it comes time to regulate, you have lots of allies in lots of places. You could generally teach people the fundamentals of why cooperation is good and why everyone dying is bad."

Blake Richards on Why he is Skeptical of Existential Risk from AI

This is a linkpost for https://theinsideview.ai/blake

I have recently interviewed Blake Richards, an Assistant Professor in the Montreal Neurological Institute and the School of Computer Science at McGill University and a Core Faculty Member at MiLA. Below you will find some quotes summarizing his takes on AGI.

Blake is not really concerned about existential risk from AI. Like Yann LeCun, he finds that AGI is not a coherent concept, and that it would be impossible for an AI to be truly general (even if we restrict the no free lunch theorem to economically valuable tasks).

Why I Interviewed Blake

Although I do not agree with everything he says, I think there is value in trying to interact with AI researchers outside of the AI Alignment bubble, understanding exactly what arguments they buy and do not buy, eventually nailing down some cruxes that would convince them that AI existential risk is worth thinking about.

Better understanding LeCun's position has been valuable for many on LessWrong (see for instance the 2019 debate with Bengio and Russell), and Blake thinking is close to Yann's, given they are part of a similar philosophical bent.

Why you Might Want to Talk to Skeptics

Another exercise I found insightful was (mostly incorrectly) assessing people's views on AI Alignment and AI timelines, which made me understand better (thanks Cunningham's law!) the views of optimists (they turned out to be pretty close to Richard Ngo's reasons for optimism at 11:36 here).

In any case, I recommend to people who are in touch with ML researchers or practitioners to 1) get to a level where they feel comfortable steelmanning them 2) do a write-up of their positions on LW/EAF. That would help nail down the community's understanding of what arguments are convincing or not, and what would make them change their mind.

To that end, here are what Blake has to say about his position on AGI and what could make his change his mind about existential risk.

Generalizing to "All Sort of Tasks We Might Want It To do"

"We know from the no free lunch theorem that you cannot have a learning algorithm that outperforms all other learning algorithms across all tasks. [...] Because the set of all possible tasks will include some really bizarre stuff that we certainly don't need our AI systems to do. And in that case, we can ask, "Well, might there be a system that is good at all the sorts of tasks that we might want it to do?" Here, we don't have a mathematical proof, but again, I suspect Yann's intuition is similar to mine, which is that you could have systems that are good at a remarkably wide range of things, but it's not going to cover everything you could possibly hope to do with AI or want to do with AI."

Contra Transfer Learning from Scaling

"What's happened with scaling laws is that we've seen really impressive ability to transfer to related tasks. So if you train a large language model, it can transfer to a whole bunch of language-related stuff, very impressively. And there's been some funny work that shows that it can even transfer to some out-of-domain stuff a bit, but there hasn't been any convincing demonstration that it transfers to anything you want. And in fact, I think that the recent paper... The Gato paper from DeepMind actually shows, if you look at their data, that they're still getting better transfer effects if you train in domain than if you train across all possible tasks."

On Recursive Self-Improvement

"Per this specificity argument, my intuition is that an AI that is good at writing AI code might not have other types of intelligence. And so this is where I'm less concerned about the singularity because if I have an AI system that's really good at coding, I'm not convinced that it's going to be good at other things. [...] Instead, what I can imagine is that you have an AI that's really good at writing code, it generates other AI that might be good at other things. And if it generates another AI that's really good at code, that new one is just going to be that: an AI that's good at writing code."

Scaling is "Something" You Need

"Will scale be literally all you need? No, I don't think so. In so far as... I think that right off the bat, in addition to scale, you're going to need careful consideration of the data that you train it on. And you're never going to be able to escape that. So human-like decisions on the data you need is something you cannot put aside totally. But the other thing is, I suspect that architecture is going to matter in the long run.

I think we're going to find that systems that have appropriate architectures for solving particular types of problems will again outperform those that don't have the appropriate architectures for those problems. [...] my personal bet is that we will find new ways of doing transformers or self-attention plus other stuff that again makes a big step change in our capabilities."

On the Bitter Lessons being Half True

> "For RL meta-learning systems have yet to outperform other systems that are trained specifically using model-free components. [...] a lot of the current models are based on diffusion stuff, not just bigger transformers. If you didn't have diffusion models and you didn't have transformers, both of which were invented in the last five years, you wouldn't have GPT-3 or DALL-E. And so I think it's silly to say that scale was the only thing that was necessary because that's just clearly not true."

On the Difficulty of Long-term Credit Assignment

> "One of the questions that I already alluded to earlier is the issue of really long-term credit assignment. So, if you take an action and then the outcome of that action is felt a month later, how do you connect that? How do you make the connection to those things? Current AI systems can't do that."

> "the reason Montezuma's revenge was so difficult for standard RL algorithms is, if you just do random exploration in Montezuma's revenge, it's garbage, you die constantly. Because there's all sorts of ways to die. And so you can't take that approach. You need to basically take that approach of like, "Okay up to here is good. Let's explore from this point on." Which is basically what Uber developed."

On What Would Make him Change his Mind

> "I suppose what would change my mind on this is, if we saw that with increasing scale, but not radically changing the way that we train the... Like the data we train them on or the architectures we use. And I even want to take out the word radically without changing the architectures or the way we feed data. And if what we saw were systems that really... You couldn't find weird behaviors, no matter how hard you tried. It always seemed to be doing intelligent things. Then I would really buy it. I think what's interesting about the existing systems, is they're very impressive and it's pretty crazy what they can do, but it doesn't take that much probing to also find weird silly behaviors still. Now maybe those silly behaviors will disappear in another couple orders of magnitude in which case I will probably take a step back and go, "Well, maybe scale is all you need"."

*(**disclaimer for commenters**: even if you disagree about the reasoning, remember that those are just intuitions from a podcast whose sole purpose is to inform about why ML researchers are not really concerned about existential risk from AI).*

Victoria Krakovna on AGI Ruin, The Sharp Left Turn and Paradigms of AI Alignment

This is a linkpost for https://www.theinsideview.ai/victoria

Victoria Krakovna is a Research Scientist at DeepMind working on AGI safety and a co-founder of the Future of Life Institute, a non-profit organization working to mitigate technological risks to humanity and increase the chances of a positive future.

In this interview we discuss three of her recent LW posts, namely DeepMind Alignment Team Opinions On AGI Ruin Arguments, Refining The Sharp Left Turn Threat Model and Paradigms of AI Alignment.

This conversation presents Victoria's personal views and does not represent the views of DeepMind as a whole.

Below are some highlighted quotes from our conversation (available on Youtube, Spotify, Google Podcast, Apple Podcast). For the full context for each of these quotes, you can find the accompanying transcript.

The intelligence threshold for planning to take over the world isn't low

> Michaël: "Do you mostly agree that the AI will have the kind of plans to disempower humanity in its training data, or does that require generalization?"
>
> Victoria: "**I don't think that the internet has a lot of particularly effective plans to disempower humanity.** I think it's not that easy to come up with a plan that actually works. I think coming up with a plan that gets past the defenses of human society requires thinking differently from humans. I would expect there would need to be generalization from the kind of things people come up with when they're thinking about how an AI might take over the world and something that would actually work. Maybe one analogy here is how, for example, AlphaGo had to generalize in order to come up with Move 37, which no humans have thought of before. [...]
>
> The same capabilities that give us probably creative and interesting solutions to problems that, like Move 37, could also produce really undesirable creative solutions to problems that we wouldn't want the AI to solve. I think that's one argument that I think is also on the AGI Ruin list that I would largely agree with, that **it's hard to turn off the ability to come up with undesirable creative solutions without also turning off the ability to generally solve problems that we one day want AI to solve**. For example, if we want the AI to be able to, for example, cure cancer or solve various coordination problems among humans

and so on, then a lot of the capabilities that would come with that could also lead to bad outcomes if the system is not aligned." (full context)

Why Refine The Sharp Left Turn Threat Model

(On the motivations for writing Refining The Sharp Left Turn Threat Model , a Lesswrong post distilling the claims in the sharp left turn thread model as described in Nate Soares' post)

> "Part of the reason that I wrote a kind of distillation of the threat model or a summary how we understand it is that I think the original threat model seems a bit vague or it wasn't very clear exactly what claims it's making. It sounds kind of concerning, but we weren't sure how to interpret it. And then when we were talking about it with within the team, then people seem to be interpreting it differently. It just seemed useful to kind of arrive at a more precise consensus view of what this threat model actually is and what implications does it have. Because **if we decide that the sharp left turn is sufficiently likely, that we would want our research to be more directed towards overcoming and dealing with the sharp left turn scenario.** That implies maybe different things to focus on. It's one thing that I was wondering about. To what extent do we agree that this is one of the most important problems to solve and what the implications actually are in particular. [...]
>
> **The first claim is that you get this rapid phase transition in capabilities**, rather than, for example, very gradually improving capabilities in a way that the system is always similar to the previous version of itself. **The second claim is that assuming that such a phase transition happens, our Alignment techniques will stop working. The third claim is that humans would not be able to effectively intervene on this process.** For example, like detecting a sharp left turn is about to happen and stopping the training of this particular system or maybe coordinating to develop some kind of safety standards or just noticing warning shots and learning from them and so on. These are all kind of different ingredients for a concerning scenario there. Something that we also spend some time thinking about is what could a mechanism for a sharp left turn actually look like? What would need to happen within a system for that kind of scenario to unfold? Because that was also kind of missing from the original threat model. It was just kind of pointing to this analogy with human evolution. But it wasn't actually clear how will this actually work for an actual Machine Learning system." (full context)

A Pivotal Act Seems Like A Very Risky And Bad Idea

> "**There's this whole idea that in order to save the world, you have to perform a pivotal act where a pivotal act is some kind of intervention where you prevent anyone in the world from launching an unaligned AGI system**. I think MIRI in general believe that you can only do that by deploying your own AGI. Of course if you are trying to deploy a system to prevent anyone

else from deploying an AGI, that's actually a pretty dangerous thing to do. **That's one thing that people, at least in our team, disagreed with the most. The whole idea that you might want to do this or, not to mention that you would need to do this, because it just generally seems like a very risky and bad idea.** The framing bakes in the assumption that there's no other way to avoid unaligned AGI being deployed by other actors. This assumption relies on some of MIRI's pessimism about being able to coordinate to slow down or develop safety standards. [...]

I do feel somewhat more optimistic about cooperation in general. Especially within the West, between western AI companies, it seems possible and definitely worth trying. Global cooperation is more difficult, but that may or may not be necessary. But also, **both myself and others on the team would object to the whole framing of a pivotal act as opposed to just doing things that you would need that increase the chances that an unaligned AGI system is not deployed. That includes cooperation. That includes continuing to work on Alignment research, continuous progress** as opposed to focusing on this very specific scenario where some small group of actors would take some kind of unilateral action to try to stop unaligned AGI from being deployed." (full context)

Ethan Caballero on Private Scaling Progress

This is a linkpost for https://theinsideview.github.io/ethan

Some quotes from the latest episode of my podcast, The Inside View. You can access the audio, video and transcript here. The key insight is that we are only seeing the tip of the iceberg w.r.t. Large Language Models Scaling, and Alignment can be seen as an inverse scaling problem.

Alignment as an Inverse Scaling Problem

> "All alignment is inverse scaling problems. It's all downstream inverse scaling problems. All of alignment is stuff that doesn't improve monotonically as compute, data and parameters increase [...] because sometimes there's certain things where it improves for a while, but then at a certain point, it gets worse. So interpretability and controllability are the two kind of thought experiment things where you could imagine they get more interpretable and more controllable for a long time until they get superintelligent. At that point, they're less interpretable and less controllable."
>
> "Then the hard problem though is measurement and finding out what are the downstream evaluations because say you got some fancy deceptive AI that wants to do a treacherous turn or whatever. How do you even find the downstream evaluations to know whether it's gonna try to deceive you? Because when I say, it's all a downstream scaling problem, that assumes you have the downstream test, the downstream thing that you're evaluating it on. But if it's some weird deceptive thing, it's hard to even find what's the downstream thing to evaluate it on to know whether it's trying deceive. "

On Private Research at Google, Deepmind

> "I know a bunch of people at Google said, yeah, we have language models that are way bigger than GPT-3, but we just don't put them in papers. "
>
> "The DeepMind language models papers, they were a year old when they finally put them out on arXiv, Gopher and Chinchilla. They had the language model finished training a year before the paper came out. "

On Thinking about the Fastest Path

> "You have to be thinking in terms of the fastest path, because there is extremely huge economic and military incentives that are selecting for the fastest path, whether you want it to be that way or not. So, you got to be thinking in terms of, what is the fastest path and then how do you minimize the alignment tax on that fastest path. Because the fastest path is the way it's probably gonna happen no matter what."

"The person who wins AGI is whoever has the best funding model for supercomputers. Whoever has the best funding model for supercomputers wins. You have to assume all entities have the nerve, 'we're gonna do the biggest training run ever', but then given that's your pre-filter, then it's just whoever has the best funding models for supercomputers. "

On the funding of Large Language Models

"A zillion Googlers have left Google to start large language model startups. There's literally three large language model startups by ex-Googlers now [1]. OpenAI is a small actor in this now because there's multiple large language model startups founded by ex-Googlers that all were founded in the last six months. There's a zillion VCs throwing money at large language model startups right now. The funniest thing, Leo Gao, he's like: 'we need more large language model startups because the more startups we have, then it splits up all the funding so no organization can have all the funding to get the really big supercomputer [...] they were famous people like the founder of the DeepMind scaling team. Another one is the inventor of the Transformer. Another one was founded by a different person on the Transformer paper. In some ways, they have more clout than like OpenAI had. "

1. ^

   adept.ai, character.ai, and inflection.ai.

Evan Hubinger on Homogeneity in Takeoff Speeds, Learned Optimization and Interpretability

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual.

Below is the transcript of my chat with Evan Hubinger, interviewed in the context of the inside view a podcast about AI Alignment. The links below will redirect you to corresponding timestamps in the youtube video.

Outline:

- Evan's background @ MIRI & OpenAI
- Coconut & functional programming (repo)
- Homogeneity in AI takeoff (AF post: homogeneity in AI takeoff scenarios)
- Reproducing SoTA & openness in multipolar scenarios
- Quantilizers & operationalizing strategy stealing (AF post: operationalizing compatibility with strategy stealing)
- Risks from learned optimization & evolution (paper)
- Learned optimization in Machine Learning (green arrow env, chest/key env)
- Clarifying Inner AI Alignment terminology (AF post: clarifying inner alignment terminology)
- Transparency & Interpretability (AF post: Chris Olah's view on AGI Safety)
- 11 proposals for safe advanced AI (AF post: an overview of 11 proposals for building safe advanced AI)
- Underappreciated problems in AI Alignment & surprising advances in AI

Michael: To give a bit of context to the viewers, MIRI is the Machine Intelligence Research Institute, can you just like give a brief recap on like how long have you been working there and what you do there and what have you been doing before, like in the past few years.

Evan: Yeah, so I work at MIRI, I'm a research fellow there, I work on broadly, to sort of very broadly, I think about inner alignment, which is sort of the problem of how do we align the models that we train with the sort of objectives that we're trying to train them on? I tend to think about this sort of problem from a prosaic perspective, from the perspective of thinking about concrete machine learning systems, also from a theoretical perspective, so trying to think about machine learning systems and understand what they're doing using sort of more theoretical tools and abstract reasoning rather than sort of concrete experiments, so that sort of broadly what I work on and what I think of.

Michael: You're talking about empirical work, so I remember when I first learned about you, it was because I was at a conference in 2018 with Vladimir Mikulik and it was just after the MIRI summer fellows and you guys were writing the mesa-alignment paper. It was mostly theoretical. And then I think you worked at the OpenAI, on theory there, but the whole company was more, you know, experiment focused. And also I think maybe before you have some more like software engineering in your

background. So you have like different interests and different expertise in both domains.

Evan: Yeah, so I did do a bunch of software engineering stuff before I got into AI Safety. The biggest thing that I might be known for in that domain is that I wrote a somewhat popular functional programming language called coconut. And then I , actually the first thing I did in AI Safety was I did an internship at MIRI doing sort of more functional programming type theory stuff, but all sort of software engineering, and then sort of went to the MIRI summer fellows program, worked with Vlad and other people on the Risk from Learned Optimization paper. And then after that, when I graduated, when I finished my undergrad, I went to OpenAI, and I did some theoretical research with Paul Christiano there. And then when that was done, I joined MIRI as a full-time researcher. And that's where I have been for the past year and a bit.

Michael: For people not familiar with AI Alignment, which I think is not the most of the listeners, Paul Christiano was one of the OG in empirical AI Alignment research now after Yudkowsky. So interning with him is pretty, pretty high bar and it's pretty good to have done that after your undergrad and, yes, so the library built was also used as a function of programming and the stuff at MIRI was also functional programming. So if I remember MIRI has had one of the leading programmer in functional programming mostly on Haskell, maybe I'm wrong.

Evan: Are you talking about Ed Kmett?

Michael: Yes.

Evan: Yep, he works at MIRI. He is a really big Haskell guy.

Michael: Coconut is more like an interpreter or something on top of python.

Evan: It compiles to python. And the syntax is a superset of Python three and then it compiles to any Python version and also lots of functional features and stuff.

Evan: It compiles to Python source.

Michael: OK, python source. So is it like very poorly optimized? If you need to, like, put something that converts to python code then it would be like super slow, maybe.

Evan: Not super, it's like the same speed as Python because you just compile it to Python and then you run the Python.

Michael: Ok, python source. When you're compiling, you're not doing perfect work.

Evan: But you don't you don't compile at run time. Like with C you're not like, well, the speed of C is, first you have to compile it and compiling it takes a really long time. Then you have to link it and linking takes a really long time and then you have to run it. You're like no the speed of C is you compile it beforehand, and then you check how long it takes to run.

Michael: OK, well, I get it now. Um, yeah, and I think that's especially interesting to me because I think there's a lack in open source, at least in the AI Alignment place. So even if coconut is not especially for AI Alignment, I think functional programming might be useful for MIRI at some point.

Evan: Yeah, it was definitely how I first got into like doing stuff at MIRI because MIRI was like doing a lot of commercial programming stuff and I had a strong functional programming background and they were like, you should come do some stuff here.

Michael: I think at some point they were hiring more programmers I don't know if that's still the case. They're still hiring more programmers?

Evan: Yeah, I think that has changed, I don't know what exactly the current status is, I'm not the person to talk to you about that.

Michael: OK, no problem. Yeah. So I guess now your job is mostly writing interesting research on the AI Alignment forum, something posting it on arXiv. And yeah, I think your posts are both very precise in terms of vocabulary terminology and clear and also short. So you can read one and not spend too long and then you understood most of the points. That's a good thing. Most people don't try to distill what they think about, and you also try to give concrete proposals for how to solve this, which is kind of a shift I've seen in the past three or four years with people having their concrete agendas on how to solve things. So one contrarian view, not contrarian but some view you had that was opposite to most of the AI Alignment forum, which is a forum for talking about AI Alignment, is most people were talking about, take off speeds like how fast, to clarify what Paul Christiano meant by fast, what Bostrom meant by fast, slow, take off, and then you mentioned something else, which was homogeneous versus heterogenous takeoff. So maybe you can talk a bit about that, like summarize a bit the blog.

Evan: There's a lot of discussion where people talk about takeoff speeds, about fast versus slow takeoff, continuous takeoff versus discontinuous takeoff.

Michael: You can even, like, summarize takeoff. Takeoff is maybe a bit poorly defined. You can even define that if you want.

Evan: When people talk about AI takeoff they're sort of talking about, well, at some point we're going to build really powerful AI systems and then something happens after that. What does that look like? Do we build the first really powerful AI system and then it very quickly dominates the world, it has a decisive strategic advantage, the only thing that really matters is what that system does, or we build the first AI system and it's really powerful and transformative, but then we build another system and we build another system, and there's a lot of different systems, but also they exist in the world the same time.

Michael: Multipolar scenarios.

Evan: Yes, unipolar versus multipolar. There's a lot of different things you can talk about, So how quickly do the systems get better? Are they sort of big discontinuities in how they get better. How concentrated is power over these systems, et cetera, et cetera. One thing that I have sort of talked about in the past in this regard is this homogeneity idea, which is, I guess in my eyes, the axis that I care about most, it feels like the most relevant and also the one that I feel like the more confident in and I can make more definitive statements aboutm, where homogeneity is saying: a homogeneous takeoff is one where all of the AI's are basically equivalently aligned and an inhomogeneous takeoff is one where we have a bunch of different AIs that are sort of varying degrees of aligned. So, there's lot of different things that happen in these different situations, and there's also sort of different aspects of homogeneity. So by default, I sort of mean alignment, but also we can talk about sort of homogeneity of other aspects of how the AIs are built. So I expect quite a lot of homogeneity I think

by default. I expect to sort of be in a situation where we have a lot of different AIs running around, but all of those AIs are basically all just kind of copies of each other or like very similar. Or in a situation where we just have a very small number of AIs. But they're still just like. If you have only one AI, then it's sort of homogeneous by definition. And so I think this is like in some sense the more important dimension, so I think a lot of times when people talk about sort of this fast, really fast takeoff scenario means that we have to get that first AI system totally correct because if we don't get the first AI system totally correct in a really fast takeoff scenario it very quickly controls all the resources, and sort of the only thing that matters is that system, whereas is in a sort of slow takeoff scenario we get the one system but then there's a lot of other systems that are competing for power and resources and we sort of have the opportunity to intervene and sort of control things more as it sort of continues to develop. And my take is something like, I don't think even in the second scenario, that we actually have the ability to really do much, even if there's lots of different AI systems running around competing for resources after the point at which we build the first powerful advanced AI system, given that I expect all of those other AI systems to basically just be copies of the first system because if they're all just kind of copies of each other then what really matters is did we align that first system properly? And so do we have a bunch of systems running around that are all basically aligned or do we have a bunch of systems running around that are all basically misaligned. And so, therefore, I'm like, well, if you believe in homogeneity, then that basically means you have to believe that the sort of first powerful, advanced system that we build is really important and critical and like aligning it is the most important thing, regardless of whether you believe we're actually going to end up in a very fast takeoff scenario. And so then there's the question of why do I believe in homogeneity? So I think basically I believe in homogeneity for a couple of reasons. First is that I think I expect pretty strongly that we'll be in a regime where the cost of training your AI is just like much, much, much higher than the cost of running it. And this creates a bunch of sort of particular economic incentives. So like one thing that it does is it means that you would generally rather use somebody else's AI whenever possible than to have to train your own one. And also in the situations where you do have to train your own AI, because let's say, for example, you don't want to use your, like, geopolitical rivals' AI or whatever, then, you're probably going to do so very conservatively, like if you have to spend a trillion dollars to train your, like, one AI system just because you don't want to use their AI system because you don't trust, like, you're the US and you don't trust China's AI, then you're going to spend a trillion dollars pretty conservatively. You're just going to be like, well, we basically kind of know what they did. Let's just do the same thing because we really don't want to spend a trillion dollars and have it not work. In this sort of regime, I expect, I think there's some other assumption here, which is like, well, I think basically if you're like running essentially the same training procedure, then you get essentially the same degree of alignment. And like the really small piddly details are not what what sort of contributes very much to like whether your thing is aligned. It's mostly like, do you basically get the incentives and biases right or not.

Michael: Right. So I guess your example with the US and China would be something like GPT-3 taking $5M or $20M to train, maybe much more to like pay the salaries and the experiments that went beforehand. This is like relatively cheap for a government now. But if we're talking about billions of dollars, then maybe it's like more expensive or trillions for an entire country.

Evan: So we get to the point where it's like a substantial fraction of the entire GDP of your country. You really don't want to like be in a situation where it's like we're going to spend one percent of our GDP, like there's this other country that has already done

this, like built a really powerful AI and has demonstrated that it works in this particular way, like, let's say they use Transformers or whatever. Like you really don't want to then spend one percent of your entire GDP trying to make it work with LSTMs. You'll just go like "no they did it with Transformers, we're gonna do the same thing. Like we just want to have our own version of what they did." And so my default expectation is this sort of conservatism, which is like, well, probably people are just gonna copy what other people are doing. And so, like, it really matters what the first thing is that's like super impressive enough that it gets everyone to copy it.

Michael: Sure. So if we take the example of GPT-3, then they didn't release the weights and it was super expensive to reproduce, like according to different sources, Google might have reproduced it pretty quickly, but it's not public information. Then there's like Eleuther.ai who tried to reproduce for months. And then after like six months or something then they produced something that had somehow similar results. I'm not sure how close it is.

Evan: Then what they're aiming for is basically a reproduction. It's not worth spending all of those resources if you don't already have the evidence that it's like going to succeed.

Michael: They know it's going to succeed, there's a paper where it's known to succeed. But at least you have the architecture and, you know, the outputs, the loss, you know what's the expected behavior. But you're surely not sharing the weights nor the data. So you have to both like scrap the data one the internet and then get the like source code.

Evan: But still the data collection procedure was pretty similar, right. Like we're just like, well, the data collection procedure, the basic data collection procedure is we're just going to scrape a bunch of like large, you know, swathe of Internet data. You know, maybe they didn't explicitly release the data. But I think that, like, I guess my take is that if you're basically, if you have essentially the same data collection procedure, you have, like essentially the same architecture, you have essentially the same training process, then you're basically going to get the same alignment properties out the other end. And some people would disagree with that. And we can talk about it why they would believe that.

Michael: I think I broadly agree with that. I might just be pointing at, maybe in the future when when we get closer and closer to some kind of human-level AI then maybe people might share less of the research about, like how they collect the data and stuff, and we just have like super-hard-to-replicate results because we don't even have the architecture or something. We just have to output or, hum. I'm not sure if the OpenAI in 2025 or 2030 will actually share everything with the other companies and how long it will take to produce it will depend... [Maybe they will be] a big enough gap that they can have a comparative advantage that leads them to lead the race or something.

Evan: It's hard to keep your just like basic methods secret. I think I mean, so currently, most of these companies don't even try. Right. Like, you know, like you're saying, you know, OpenAI I just publish papers on exactly what they did. And they don't publish the weights but they publish the paper saying what they did. Anything with DeepMind, Google, basically. And I think a part of the reason for this is that even if you wanted to keep it secret, it's pretty hard to keep just like these very general ideas secret because like one person leaves and then like, you know, they can explain

the general ideas. They don't need to steal anything from you. It's just like they already have the idea and the idea is very general.

Michael: The moment you hear about it, it's like the moment we heard about, when GPT-3 became mainstream, let's say July, it was released in maybe May, and maybe they had the results maybe in January or something. And they had some close results and then they had to like, make it better, improve it for publishing it to NeurIPS or something. For Dall-E maybe they knew about multimodal for a while and at DeepMind I think the politics is that they hold their private research, a bunch of research is going on for like months and then they try to publish it to Nature neuroscience, or like Nature. And so you have all those like deadline for papers where they have those those six months advantage or maybe a year advantage where they all have private information that they don't share to other companies.

Evan: Yeah, there's definitely a delay. But like at the point where you have this really powerful system and you start doing things with it such that other people start wanting to build their own really powerful systems. I expect convergence.

Michael: Convergence of people sharing stuff?

Evan: Like people are going to figure out what you're doing, even if you don't try to share things like it's just too hard to keep it secret when you're like you're like having this big model, you're out and you're doing all of the stuff people are going to figure out, like what the basic process that lead you to...

Michael: Like reverse engineering or social hacking? Like you just need one person to basically describe the idea. Like, honestly, I expect, it's so insecure with like, I don't know, like, can you imagine trying to keep the basic idea like we used attention secret. Like, I just like I don't it's not gonna happen.

Michael: We used a transformer, but please don't tell anyone.

Evan: I think it's totally practical to keep your weight secret because if somebody wants to steal your weights, they have to actually copy them and exfiltrate the data. And like, it's clearly illegal. And like, you know, it might still happen. Like, I totally I think it's still quite possible that, you know, you'll have, like, hacking people trying to steal your weights or whatever.

Evan: It's at least plausible that you could keep your weight secret like there's just no way to keep "we used attention to do really big language models" secret.

Michael: Hmm, I think I think one example that goes in your direction is Dall-E. This github account lucidrain's reproduced Dall-E in a couple of weeks I think, maybe like less than two months, something like that. I think, like Dall-E was maybe end of December beginning of January and lucidrains published it in maybe February or something. So for some experiments like it gets faster and faster to reproduce it. I was talking to someone else at OpenAI who told me that he expects multipolar scenarios to be like kind of the default because as an entire community, we get better and better at reproducing what the best labs do. So the time between something GPT-3 and reproduction gets closer over time. I guess one counterargument or question I had about this blog was, you say that when the first is alined and people will copy it and by default it will be aligned because it has the same architecture, but imagine the thing that is aligned as like some kind of those laws of robotics or, oh, don't kill people or be aligned with human values and stuff. If you're trying to be adversarial and try to beat the first AI or smartest AI alive, not implementing this aligned features, you could

just like be very adversarial and attack the first one who would not attack you back because he's like human aligned right.

Evan: One thing which you could imagine is a situation where, like the first person, the first organization or whatever builds an AI, they successfully make it aligned and then somebody else decides to build an AI and not do the alignment part. This relies on there being some separability between the alignment parts and the nonalignment parts, which is unclear. I'm not sure if I actually expect there to be this sort of separatability, but then also it relies on you not caring about alignment as a good desideratum, but that seems really unlikely. Alignment is pretty useful for almost anything you want to use your AI for, like you would really like your AI to do the things you wanted to do. It doesn't really seem much different than just when we're in a situation which is like what I expect where these sort of training runs have a truly astronomical cost, right. Where like your little research lab in like a university or whatever isn't going to be able to replicate in the biggest lines because the thing costs like, you know, billions of dollars for a single training run, trillions of dollars or whatever. Then you're in a situation where, like, you really don't want to risk spending a billion, a trillion dollars or whatever and have it not be aligned, have it not do the thing which you want it to do. You're going to copy whatever alignment features the other team did that they successfully demonstrated. It might be the case that, in fact, the first thing only looks like, maybe it's not really aligned. You're still going to copy it, I think. And so this is why I sort of have this view, which is like, well, even in a sort of very multipolar, continuous slow take off world, we still basically just care about making sure that the first really powerful system is aligned in the traditional conventional way that we think about in the fast discontinuous takeoff world.

Michael: One of the post when you're trying to clarify alignment terms, I think Paul Christiano tried to clarify it in one of his blogs, on medium, which is alignment is basically I doing what you tell him to do. So very basic terms. And then you distinguish something like impact alignment and intent alignment, which is trying to do what the human wants you to do and then impact alignment, which is not destroying the universe and not causing harm to humans. Is it essentially correct, or do you want to maybe nuance it?

Evan: Yeah. So impact alignment is in fact not doing bad things. Intent alignment is like trying to not do bad things, in my definition, and then we can further split each of these things down, so intent alignment can be split into, outer alignment, which is like is the objective that it's trained on aligned and this objective robustness thing, which is like it is actually robust according to that objective across sort of different possible situations.

Evan: The whole post has a sort of breakdown where it sort of looks at, ok how do we sort of breakdown these different ways of sort of thinking about alignment into different sub-problems.

Michael: Yeah, I was trying to find the actual actual picture from the blogpost because it's pretty good, but yeah, I think it is in clarifying Alignment terminology. I was just trying to see if I could make it as my background for fun. But I think it's pretty hard, let's see. Yeah. So I think. So what you were mentioning before is kind of companies would want something like intent alignment. Or like something that does what the Chinese government wants them to do and if the Chinese government wants to kill everyone in the US, or like take over the world, then it must be intent aligned in the sense of trying to optimize for like do the same thing that the Chinese government wants them to do. But it doesn't mean that it won't kill other countries, right? The

second actor just might want just to have something useful without it being beneficial for the entire human race?

Evan: Yes. So you still have even even if you have like you know, you solve the sort of like intent alignment style problem, and even if you have a very homogeneous world, you still have situations where you have like standard human coordination problems where you have to be able to coordinate between one, you know, one human organization is trying to, you know, fight some other human organization, and then the hope is that human coordination is a problem like, you know, human organizations fighting other human organizations is a problem that we can mostly solve via like normal social and political dynamics.

Evan: It's still hard. We did make it through the Cold War without any nuclear bombs being dropped, but it was close. Hopefully, if we're in a similar situation, we can make it through something like that again. The real problem that would be, you know, at least we want the opportunity to be able to do so, right? We want humans to be the ones that are in control of the levers of power such that it is even possible for humanity to be able to coordinate, to make it through a, like, similar to the Cold War style situation. If humans don't have access to those levers of power at all, then we can't even do that.

Michael: It's a necessary condition to have a peaceful outcome, not a sufficient one? I think that's essentially right. I agree with it. And then I guess some people might say that, the political problem is maybe the hardest one. And like, if we have some kind of very authoritarian regime that I don't know, say, is working on AIs, you can go into trouble with your work on AI, too advanced AI. And like everyone is doing good old manufacturing job or like agriculture jobs. Then if we solve the political stuff before and we have some large peace on [inaudible] government, then we solve kind of, we have more time for AI safety. That's a bit like the steelman of the other position.

Evan: Yeah, I mean, certainly I think there's lots of ways which you can approach this problem, or just like AI existential risk in general. There are like social and political aspects of things that are worth solving, as well as technical aspects. I think that currently I'm like, well, I feel like most of the existential risk comes from we're just not going to be able to solve the technical problem.

Michael: Most researchers working on this, I guess, think similar things, which I think is pretty tractable. I think the solutions you give in your other blogposts seem pretty tractable. Solving politics or human coordination or fighting against mullock seems a bit harder. I agree that it's worth thinking about it. I think we covered most of this blogpost, and then, sure, you had another one which was interesting to me because I worked a bit on AI Alignment research myself and did some open sourcing as well, on quantilizers. And so that's a bit what I'm familiar with, in terms of research. And you worked on a post on, like how to... Quantilizers is essentially if I were to summarize it would be "we try to have AIs that perform in some kind of human way without it being too bad at the task". So if you have a human demonstrating a task, let's say a robotic task or some game playing or feeding another baby or something, you want the AI not to find the optimal action because the optimal action would kind of hack the game. But you also want it to perform well and not do stupid things. So the human is like drinking alcohol a tenth of the time of the day. You don't want the AI to drink alcohol. You wanted the AI to do the normal human stuff in the afternoon. So quantilizer is essentially taking those 10 percent, or this quantile of actions that are good, that are still human-like so when the AI tries to imitate those actions... one thing about quantilizers is doing that. I guess there are other ways of seeing quantilizers, this is

one way of looking at it. And so in your posts, you're kind of talking about what Yudkowsky's... how he defines bits of optimization, which is when you have this interval of length one, of probability mass, then if you have something that is the in the half of highest utility, then you're somehow halving the space in half. So you're having one bit of optimization or something. And then the closer you get to the optimization power, the more bits of kind of optimization you need. Yeah, maybe you can say that better and nuance it.

Evan: I guess you're referring to my post on operationalising compatibility with strategy stealing.

Michael: Exactly.

Evan: In that post I talk a little bit about optimization of power and quantilizers. I give a definition of optimization power in terms of quantilizers and then I try to relate this to the strategy stealing assumption and value-neutrality verification. Maybe the best thing to do will be yes, I'm happy to talk a little about this post, I think it's certainly interesting, I wrote it. So some things which I think. Maybe I'll start with strategy stealing. I think strategy stealing is an interesting thing because I think. There's a lot of different ways to think about it. The very simple thing is that there is this... there's this formal result, which is, well, you have a game and it's symmetric, we have a bunch of different players and they're all sort of playing, you know, they they have, you know, access to essentially the same actions and stuff. Then you can always just sort of copy the strategies of any other players such that you can never sort of do worse in expectation. And if there are N players getting 1/N of the sort of total resources. You know, even if there is like... in any situation, as long as it's symmetric, you can sort of make this work.

Evan: What does this mean? Well, so one of the ways in which we can get interpretations of this and Paul sorts of talk a lot about this, is that we can sort of think about this as giving us a sort of general desideratum for AI safety, because we are currently in a situation where humans sort of control most of the resources. And so we should expect that, you know, by default, that sort of distribution of resources shouldn't change over time because we can just sort of, you know, if other agents are introduced and they start with very little resources, then we can just sort of steal whatever strategy they have. And in expectation, we should have the same proportion of total resources that we have right now and at any future point in time. But then this fails for AI, because AI has this sort of property that it might be systematically biased towards certain values over other values. So, for example, values which are very easy to specify. And so we can build reward functions but not really easily, we [=the easy to specify ones] will win out systematically, and so this breaks strategy stealing because now it's no longer symmetric because some people's values that are easy to specify will systematically [inaudible]. Similarly values that are really simple, such that they're really easy to find by default in sort of training processes will sort of systematically [inaudible]. One way in which we can think about one thing you might want a sort of aligned training process with you is not the sort of systematically better for some of these sorts of values than other values. And in particular, not be systematically better for like simple or easy to specify things than for like actual human values that we care about. And so one way in which we can define that notion is using sort of this concept of optimization power and asking, you know, to what extent is it applying more optimization power? Is it sort of able to apply more optimization power to some sort of tasks than other tasks? And in particular, if it's able to apply more optimization power to... I have this example, we consider Sundar Pichai, who is the CEO of Google, and he wants a bunch of different things. He wants like to be happy and seek for his family to

be happy, but also he wants Google to make money. And so he has like a really powerful AI system and it's like trying to do what he wants. And so he's like, ok, you know, here are some things I want you to do. I want you to, like, you know, find some ways to make Google money and also, you know, find some ways to like help me, you know, figure out my life. And also, he probably cares about humanity and want humanity to be in a good spot overall, but also he wants to make money obviously. And so this AI goes out and tries to do this. But there's like a real problem if the AI is just like much, much, much better at making money for Google than it is at any of these other things. Because then it goes out and it's like, well, it makes a bunch of money for Google and it's really, really good at making money for Google, but it's like very bad at like doing any of those other things, it doesn't really know how to, like, put the world in a good spot for humanity. It doesn't really know how to make Sundar happy. He doesn't really know how to do any of these other things that Sundar cares about. And so from Sundar's perspective, what ends up happening is that, you know, some of his values lose out to all of other values. And so in the long run, we end up in a situation where we built AIs that systematically favor the development of a sort of enhancement of certain values. The enhancement of competition values, like getting more money for Google, at the expense of these other values, like, you know, is the world good? And this is bad, so we'd like to avoid this.

Michael: Right. I think that it kind of resonates with how easy it is to hack the humans brain and optimize for like Facebook ads or TikTok views and it's harder to specify make humans happy in the long term. So like, we would kind of converge towards easy-to-hack-brains behaviors and maybe like even like optimizing for the crypto market or optimizing for the trading market is something with very little information and very little dimension compared to like visual inputs. So maybe AI would be good at things that are easy to do now and that are tractable in terms of input space, I guess. But then for Sundar, what you're saying... AI would converge to what is easy and if what is easy is maximizing profit then it will do that instead of other things. But if it understood that what Sundar wanted was actually Google making money to benefit the world and making his life good, it will not create bad Google doing bad for humanity and having Sunar work overnights and like not spend time with the family or something.

Evan: That's unclear, right? You can imagine a situation that is just kind of like the current world, but where like we know how to build AI systems that do like very simple tasks that we can specify, we don't know how to do systems that are really complex, hard to specify tasks then like we could very easily end up in a situation where due to competitive pressure, Sundar is just kind of forced to use these systems and sacrifice all these other values to make sure that Google is able to make money. But like, there's just no ability to because the only powerful actors in the world are these AIs but they can only be used for these sort of simple tasks, then you're forced competitively to keep deferring to them and giving them more power resources to be able to give you more power and resources that you never get to a point where you can actually use that power and resources for what you actually care about. To be clear this isn't the sort of world that I expect by default, but it's it's worth sort of pointing out as like, in a sort of way of thinking about a particular type of alignment problem that is not the traditional alignment problem and doesn't necessarily, isn't necessarily solved, even if you solve more sort of other aspects of the alignment problem.

Michael: Interesting. So, yeah, if you don't solve other problems, then you might end up here. And I guess the thing is some kind of Google... like I see it as a very powerful Siri or Google home, where it would be like a good oracle, like Sundar Pichai coming

home and asking his Google home "What's the best strategy for tomorrow?" I guess somehow it's not that far away. Maybe like strategy-wise, running a company like Google is hard, but like chatbots that you can talk to and like ask for simple decisions. I don't know. And the link was the kind of optimisation thing is that?

Evan: Mathematically, we can use optimization power to give a definition of this.

Michael: I think that's interesting because like in the past, ok, in the first two episodes I had, the thing I called Connor's rule because Connor Leahy, he had other podcasts, on like machine learning street talks, with Yanick Kilcher and stuff, and they went on defining multiple times intelligence. The rule is like you shouldn't talk about intelligence, you should talk about optimization, or like other stuff that the AI would do and not like talk about words. So I feel like optimization is a good word and you give a bunch of different useful terminology in risk from learned optimization, you kind of introduce mesa-optimisation amongst other stuff, and then you clarify it even more in "Clarifying AI Alignment terminology", which is a reference to diagram behind me.

Evan: it appears to be inverted for me, but I can't see it.

Michael: Oh it's inverted for you. Sorry, it was inverted for me, so I inverted it back. So I need to invert it back. I don't know what the camera will do at the end. I can have both. Yeah. Go ahead. I will just remove that.

Michael: What do you want me to say?

Michael: Maybe you can start with the mesa-optimization term. How do you define optimizers and what is a mesa-optimizer?

Evan: Risks from learned optimization takes a stance, which is something like optimization is the important thing to think about. It's not intelligence or agency or any of these other things. It's like optimization is the key thing, which is similar to sort of what you were describing. Certainly there's a lot of sort of discussion around this stance or on whether this is a good stance. I think it is a stance that lets you say a lot of things because optimization is like a reasonably concrete, coherent phenomenon. And so you can say a lot of stuff about optimization. And this is sort of what Risks from learned optimization tries to do is say a bunch of stuff about optimization. I can say more I'm happy to talk more generally about what is Risk From Learned Optimization, basically saying what is the inner problem, et cetera, et cetera.

Michael: I recently re-read the actual introduction... I think there's a sequanece on the AI Alignment forum from where you define... there's an introduction where you define all these concepts pretty precisely. I feel like optimizer is, maybe you said that already but, it's like searching anything that searches over a space to find the best solution according to some objective and actively search. So, for instance, there was this example from from Daniel Filan about "is a bottle cap an optomizer" because it's kind of preventing water to go away. So it's not actually optimizing for anything, but it's something that humans optimize for. And so it's a result of an optimization process from humans.

Michael: And humans are something evolution to mind for as we're optimizing for different things, like instrumental things, like having sex without making kids or or other things. And so that's maybe some kind of disagreement I have on your examples. And I think in your podcast with Daniel Filan, AXRP, you said it is kind of useful to see humans as optimizing, searching for some solution that is not directly evolution's function. So in terms of alleles chromosomes and stuff, because like some

humans don't make kids. My counterargument to that would be that even for humans that don't make kids, they're still like kind of trying to optimize for evolution's pressure, in a bad way. So imagine very good researchers. They don't care about making kids at all, but they're just very passionate about math. So they will end up producing value for the world with their math papers that will end up in, like, more GDP or more kids in the future for other humans.

Evan: Yeah, I think this is not how it works, though, right? I think it is just actually true that, like, if you really let evolution keep running, it would not select for this sort of behavior. Like evolution certainly wants some altruism, but it also doesn't want you to, like, live down the street from a sperm bank and not go there every day. right? Like, that's insane from an evolutionary perspective, regardless of what else, whatever else you're doing.

Michael: But it's still like we're like I still feel like we're trying. So. Like our our instincts, like our primal, our lizard brain still wants to optimize for evolution. It's just that we're bad at it. Or that we've evolved for like building those tribes. And society, that is a proxy for building more kids.

Evan: The keyword there is proxy. The things that we care about are in fact proxies for what evolution cares about. But they're not the same, right? Like you can certainly tell lots of stories about it. And It's true because there are proxies. You can think about status, and power and, you know, sex and all of those things could be our proxies for and they're related to in the ancestral environment, the sort of pass their genes on. But we don't actually care about passing our genes on, at least most of us don't. You know, I think, well, something like the sperm, do you wanna eat sperm or eggs is a good example of, like, you know, most people don't or would have to be paid to do it. And you know evolution would never want that. That's like clearly evolution is like "This is the most important thing you should be doing. You know, gotta be doing nothing with that". But from a human perspective, we care about the proxy. We're like, what I care about isn't actually literally just my genes in the next generation, you know, even like humans that can really care about like having children usually care about, like I want to be able to raise my children. I want to have a connection with my children, not just like I literally want more of my DNA in the next generation.

Michael: Those proxies are actually good at like making, like, more humans long term. Evolution evolved and found this new solution in search space, which is no the actual good stuff, is not just to be a lot. We actually need to be in some kind of tribes and have social defense to find dinosaurs or monkeys or something. And then if everyone was spending sperm, would give sperm to solving...

Evan: No, evolution doesn't work at a group level though, it works primarily on an individual level. And so evolution is happy to evolve to extinction on a group level because it's primarily selecting on an individual level.

Michael: Hmm. But wait, so if you're selecting genes.

Evan: This is why we have things like selfish genes. It doesn't actually help you it just like copies itself from place to place. Evolution isn't just selecting for the performance of the whole group, but it's very explicitly selecting for your individual performance. Another example of this is like sex ratios. So like in theory, you would like evolution for like the maximum production of additional children would want like significantly more females in each generation than males. But in fact, what we see is that across species, the sex ratio converges to 50 percent. And the reason that converges 50

percent is that from a selfish, individualistic perspective, even if you're in a population where there are greater than 50 percent females, then you are in an advantage passing on your genes to the next generation, if you have a male child and you're at a disadvantage if you have a female child. And so despite the fact that evolution from a group perspective would rather have a sex ratio that is not 50 percent, from an individual perspective, it has to be 50 percent because of like it's sort of the only stable equilibrium from a sort of selfish, individualistic perspective and evolution, primarily selects on the individual.

Michael: It's like a bunch of individuals with egoistic genes that converge to some Nash equilibrium at the society level.

Evan: Well, so we can also certainly talk about why is it that humans are altruistic? Where did that come from evolutionarily? I think the like leading theory is something like it's good for, it's useful for cooperation. Being altruistic is helpful for your ability to cooperate with the rest of the group because we care about the rest of the group and they care about you, that you can cooperate really seriously with them. And so in some sense, altruism is selfishly useful in this perspective. From an evolutionary perspective. It's like evolution would rather have each individual be more altruistic because it helps them work better with the group and less ostracized by the group and therefore have a more likely for that individual to have more children. And so this is a individualistic story of why, from the perspective of a single individual, evolution would rather that individual be more obvious.

Michael: And what about, like, people being like the opposite of altruistic and just like kind of defecting all the time with altruistic people? Like this would be like the better position, right?

Evan: No, the point that I'm making is that this is not the case. For evolution, for each individual, altruism serves a purpose for helping that specific individual have more altruism.

Michael: I think they're like other distinctions you make that are interesting. So just to define the basic terms again, because I think most of the listeners are not familiar with the paper. A good analogy for evolution is what we call the base objective. Maybe a neural network is an easier example,

Evan: Maybe it is better to start with neural networks and in Risks from Learned Optimization, we're really trying to ground everything on optimization. I think one of the big things that Risks from Learned Optimization does, that sort of all previous discussion didn't do is really carefully ground everything in Machine Learning.

Michael: So, let's talk about machine learning. What's interesting is when we have optimizers like Adam or stochastic gradient descent, then you're trying to change parameters theta so that you can better classify cats and dogs. At the end of the day, you change your parameters, they might end up at inference time doing something like optimization. The example for me would be something like a recurrent neural net where you do backprop through time, where you're optimizing and you're at inference time only using the latent cells or something? Some are frozen and some are not. And then you can adapt to, um, what you get at test time. And I think that was one example of a blog on LessWrong trying to reproduce mesa-optimization. Do you have better examples, maybe, of this sort of optimization.

Evan: So the classic example that I like to use to really explain sort of what's going on with Risks from Learned Optimization organization, is this maze example. We can

imagine a situation where we train a model in a bunch of small mazes, sort of randomly generated mazes, but they're all kind of small. And we put a green arrow at the end of the maze. It gets like a picture of maze, and we put a green arrow at the end to say "this is the end of the maze, you're supposed to go here". and we train a bunch in this environment, and then we deploy to a new environment which has the following properties. It has larger mazes. The green arrow is no longer the end, it is some random location inside of the maze. But we still want the agent to get to the end, we still want the agent to read the maze. Or you can flip it and you can be like, we still want to go to the green arrow and not go to the end. Either way, the point is there are a bunch of different ways in which this agent can generalize. So here's one generalization, is it just goes to the larger mazes and it doesn't know how to solve then, it just fails to solve big mazes. And I would sort of call this "its capabilities do not generalize", it didn't learn a "general purpose means solving capability". Or, it could learn a general purpose maze solving capability, its capabilities could generalize, and it could learn to go to the end of the maze, which is what we want it to do. And so it's objective going to the end of the maze, also generalises properly. But then there's another situation, which is, its capabilities generalize, it knows how to navigate the maze successfully, but its objective fails to generalize and it learns to go to green arrow rather than to go to the end. And then what's scary about this situation is that we now have a situation where we have a very capable model, that sort of general purpose optimization procedure, but it's deploying that optimization procedure for the wrong objective to get to the wrong goal, not the one we intended to get to the green arrow instead of what we wanted, which is to go to the end of the maze. And so this is really dangerous because we have a powerful, competent model which is directed in a direction we never intended. And what's happening here is that there is this unidentifiability where on the train distribution we couldn't tell whether what it was really doing was going to green arrow or going to the end. And when we deployed in an environment where these two things came apart. Now we can tell. And if it does the wrong thing, it could be really capably doing the wrong thing in a this sort of new environment. And so this is one example of a way in which a model can sort of have failed to have objective generalization. Its objective can be generalized properly. Well, its capabilities still generalize properly, which is the sort of general sort of subheading under which it is trying to address as a problem.

Michael: So to summarize, it's good at finding green arrows, but it's not good at finding the end of the maze.

Evan: That would be that would be a situation where we're like unhappy because it's very powerful and knows how to solve mazes properly, but it isn't using those capabilities for the right reason. It's not using them to the one we wanted to use for. It's using it for this other thing instead.

Michael: I feel like it's somehow similar to people who criticize GPT-3 for not understanding what it's saying, but it's just like repeating and memorizing things. You could say that GPT-3 doesn't have what we want [it] to have, which is a natural language processing or like human understanding of words and concepts, but just has memorization. He can of memorized the way of finding the green arrow without understanding the actual task we wanted him to solve. Does it fall into this same category or is it different?

Evan: And you can certainly think about it that way. I think it's like a little bit tricky to really think about like... you know, in some sense, the objective of GPT-3 is predictive accuracy on the next token. It's a little bit hard to understand, but would it actually look like to sort of generalize well or poorly according to that? I mean, it's just like if

you have an actual distribution that is similar... You know, I guess in some sense, but we only trained it on this Web text corpus, and then it was some new setting where the underlying generators of the text are different, then it might still be trying to do predictive accuracy or you might have learned a bunch of heuristics that are not particularly accuracy. What it really learned is it should try to, in any situation, output like coherent sentences or whatever, and then it's like it doesn't actually try to model the dynamics of this new setting and get good predictive accuracy. It just tries to do the simple like, well, I learned to do these sort of heuristics for how good sentences work. And so I'm just going to keep outputting that.

Michael: Right. He found those heuristics. Two things I remember from Connor's interviews, I'm not sure if he was with me or with other people when was "we don't really know the entropy of human language", of like English. We don't even know how hard the problem is. So i's very hard to say exactly how successful it is at predicting words or understanding it because we don't have a good model of what English is. And the other one, which is kind of a funny trick, is that I think it took something like one epoch or less than an epoch... it only passed through each example once. Maybe I'm wrong. Maybe it took more than one epoch, but it kind of learned to generalize from a few data just like passing... one shot learning.

Evan: I think one epoch is in fact compute optimal in most of these really big language models.

Michael: So, yeah, that was something impressive in terms of... for people who say like it's is memorizing. It's memorizing, yes, but maybe, but from one-shot learning or something.

Evan: I don't think one epoch is very meaningful here, it's just like, well, you got to see every data point in the training data. He'd seen the whole training distribution. He hasn't seen it multiple times thing is more just say "Oh, our training process just performs better when it can extract like it's sort of already extracted the information from that from that data point the first time through. And there's sort of diminishing returns and trying to run it through a second time. And so it's not compute optimal to do so."

Michael: Just running it for one epoch enough and compute optimal and otherwise you would just like lose money because you wouldn't get as much value for dollars something. I'm trying to find the post from Matthew Barnett, because he sent me the code at some point on how to do it, and I'm just trying to put it behind me, as I think I'm trying to do now. Putting stuff behind me. So give me just one second. So it's a map with treasure and chests, keys and chests. I don't know if you remember it, then maybe we can talk about it otherwise because I kind of remember some form of this environment, but maybe you also remember it.

Evan: Yeah, it's very similar to like my maze example. Where it's just like. There's a set of objectives which are indistinguishable on training and we move to deployment, and you can see that it is like this one and not the other.

Michael: Yeah. So if I remember correctly, it you would stumble on keys and then because there would be like more keys than chests, so it would open the chest without actually knowing what opening a chest is. And, then on big environments it wouldn't really know how to do it. Or you could still do it in bigger environments. Yes, something like that. So yeah, if you're a listener, I have some code for it. And there

are people with problems to demonstrate mesa-optimisation or inner alignment failure? Is it inner alignment failure?

Evan: Well, so inner alignment the way we define it sort of requires there to be optimization. We don't know, especially in the keys versus chest, where it's simple, the model probably isn't doing any real optimization internally.

Michael: The problem with calling it optimization is that we're kind of assuming some form of complexity, or some form of, you know, he's doing some thinking or some elaborate task or finding some optimum somewhere of some precise task. So I remember there was this Lesswrong paper post a paper from Deepmind about meta-learning—meta-reinforcement learning. It was like top of the Alignment Forum for a bit where they showed that it was similar to some kind of mesa-optimization. And then, like some people commented that it was basically reinforcement learning the thing was doing. It was not some kind of very special trick. It was just like an LSTM plus some RL, and at the end you [freeze] the weights and then you get some stuff that's going to adapt to environments. I guess, like researchers can always say, you know, yeah GPT-3 is not intelligent, it's just memorizing sentences or this thing is not optimization it is just like doing whatever it was trained on to do at the beginning.

Evan: But there is a truth of the matter. Like it's an empirical question so we can look inside of models if we have the good enough transparency tools and discover how do they work. Are they doing an opposition algorithm? Are they not doing it optimization algorithm? That is something I hope we can eventually do. I don't think we currently are able to quite get there, but I am hoping that we will eventually be able to actually answer these sort of questions by literally looking inside of our models.

Michael: Just to close a bit on this, I think that this terminology is super important, so I'm just going to put that back behind me one last time because I think that's useful for the listeners. So is it on the right side for you now? What we want is alignment, which is kind of what you said about impact alignment, which is AI that doesn't do bad stuff then capability robustness is (you can correct me at anytime) the ability to generalize to harder environments or out of distribution environments. Is this correct? You need it to be capable enough to generalize well.

Evan: Yeah, but it's sort of like generalize "according to what?" is the question and capability robustness just says generalize according to whatever objective you learn. It isn't saying that you actually learn the correct objective. It's just saying according to whatever objective you learned, generalize well according to that.

Michael: Right. So, yeah, you're capable of maximizing that reward, of minimizing this loss in a more general setting than from the training data.

Evan: Importantly capability robustness has no reference to the actual reward. It's not saying according to the actual reward, you generalize well, that's like robustness. Robustness in general says according to the actual reward, you generalize well. Capability robustness is the subset of robustness that says not according to the actual reward, just according to whatever utility function, internal objective, weaks... what do I call in this post sort of "a behavioral objective", which is just like the objective that actually describes what you're optimizing for in practice. Do you generalize well, according to that, which is just showing you sort of just make a bunch of mistakes and not really know what you're doing and don't have anything coherent or are you coherently trying to do something, regardless of what that thing is the correct thing?

Michael: Generalizing in doing what you were previously trying to achieve in a new setting and then intent alignment is what we said before as "doing what the human wants to do" Like if you say, bring me some tea, bring some tea, assuming he's not killing the baby between you and the tea. Objective robustness, is... your objective is robust to what? I forgot.

Evan: Maybe a useful thing also for distinguishing between capability robustness and objective robustness would be... there's like another version of this picture where I have it in terms of just robustness.

Evan: There's the version of the top, which is how I actually think about it, but then I think if you think about these things in terms of robustness a lot, then like it may be a little bit better to start with the robustness centric version—they're equivalent.

Evan: I was trying to say... so we have robustness. In the robustness centric version, we split alignment into outer alignment and robustness at the top level where outer alignment says "is the base objective like doing the right thing". And then robustness, says "does it generalize well according to the base objective" which is on a [new] distribution, does it continue to pursue the base objective and then we can split that into objective robustness and capability robustness? And then here I think the distinction between objective robustness and capability robustness is maybe a little bit easier. Previously we had the notion of the base objective, which is just like the reward function or loss function, and then we also introduced the notion of the behavioral objective, which is like what does it appear to be optimizing? And then we say it's capability robust if it's robust according to its behavioral objective. So whatever it looks like it's optimizing, it does a really good job at optimizing that no matter where it is. So it looks like if we look at its behavior in general, it looks like it's going to the green arrow . And so we can say its behavorial objective is try to go to the green arrow. That's not what we want. We want it to go to the end of the maze. But when we look at what it's doing, it's clearly trying to go to green arrow and then we can ask how good is it at going to the green arrow ? And if it's really good to go into the green arrow , then it's very capability robust, even though it's doing the wrong thing. We didn't want it to go to the green arrow. And so the other part of robustness is objective robustness, which is how closely does that behavioral objective match onto the base objective, which is the one we actually want? And then a sub-problem of objective robustness is inner alignment, which is saying "ok, but what if specifically we have a model which is an optimizer, is running an optimization process, and then therefore it has some objective that the optimization process is optimizing for which we call the mesa-objective, and then we can ask, inner alignment ask, how close is the base-objective to the mesa-objective? And then the point of both of these diagrams, the sort of overall point is that if we get both inner and outer alignment, then... this is the part that's harder to see on this version of the diagram, on the other version of that diagram, it is very clear that inner alignment and outer alignment imply intent alignment. Which is like sort of, I think, a good justification for why it makes sense to sort of split the problem into inner and outer alignment in the situation where your model is a mesa-optimizer, that is it's like doing optimization. If it's not a mesa-optimizer, then you can't split it into... you can't sort of... you don't have inner alignment as a sort of concrete phenomenon, just have objective robustness and then maybe it makes more sense to look at it from the robustness picture. But I think if you're if you're thinking mostly about mesa-optimisers, then you're like in outer alignment plus inner alignment is your intent alignment. Both pictures are equivalent, they're just two different ways of looking at the same thing.

Michael: I think for for the listeners, the kind of errors are kind of sufficient ways of achieving X. If you have robustness and outer alignment, you get alignment and you don't even need to have inner alignment. If there's no mesa-optimization going on and you just have like one optimizer process, then you can just have one optimization process being robust and outer aligned. So those are less sufficient ways of achieving alignment, not necessary ones. Is this correct?

Evan: Yeah.

Michael: I think what you were saying is interesting because I studied Inverse Reinforcement Learning (IRL), where where the goal is to... we have a human having some behavior, doing some stuff and it's trying to guess [the human's] reward function from his behavior. And [the human's] reward function is kind of what he wants to do and could be mapped to like his values or something of some sort. If the human was performing optimally according to his reward function, then from his behavior you could infer his reward function. And so this kind of behavioral objective is what an AI would be doing if it was optimizing for the human's objective function, if IRL was tractable in some way.

Evan: Right, yes, you can think of the behavioral objective as being related to IRL (Inverse Reinforcement Learning) in the sense that if you did Inverse Reinforcement Learning on some model, then you can think of the objective that you get as a result of doing that, as that model's behavioral objective,

Michael: For any sequence of actions... for any mapping from state to action, you can construct a set of optimal policies according to those possible reward functions, right? Or utility functions or reward functions. Cool, I think we covered that pretty well. Sorry for saying basic stuff, but I think I think most of the audience doesn't know about this paper anyway. Doesn't mean that it's not a very essential one and one of the most important one in the past few years. This means that my audience is not literate om that. So you talked a bit about transparency and how it's important to solve the AI Alignment problem. And, I guess, Chris Olah is an important actor in that space. There's other people I met or talked to in the Clarity space, I think of Nick Cammarata. And I think they're like, it takes a lot of time to write a good distill post, to explain this stuff well. And it's a lot of effort. And it's somehow... maybe you get less exposure than, like a tweet or something... but somehow you can say that gaining understanding of how ML models work is accelerating ML research, and he's also giving a good feedback loop between how do we align those? And I think in your post, you kind of gave both of the arguments and counter-arguments for Chris Olah's views. And you're the best proxy on, or one of the best proxy of Chris Olah's views on that today.

Evan: Yeah. So, yes, I think you're referring to the post I wrote sort of summarizing...

Michael: Chris Olah's views on AGI Safety I think.

Evan: Yeah, it was a while ago just after talking with Chris. I think, Chris, that a lot of interesting stuff to say about AI safety. And, I think it's like under... I don't know, at least at the time I felt like it was under appreciated and like, not really like, people weren't engaging with this sort of way of looking at the [interpretability problem as much as] AI Safety, as much as I wish they were. It's been a while since I talked with Chris about this stuff, so I'm not necessarily up to date on all the stuff these days.

Michael: I think it's from November 2019. So one and a half years old.

Evan: Yeah, I think it was like a reasonably accurate sort of like, I gave the graph to Chris and a bunch of times, going back and forth, trying to get like, what does he make sure he agrees with it and stuff? So I think it was like reasonably accurate, at least a good summary of sort of stuff he was thinking about then. So, yes, I think it's definitely worth... Yeah. And I definitely think Chris is sort of doing a lot of transparency stuff and is still probably the person who's been most stuff in the [explainability?] space of general transparency stuff, that is at least... that is relevant to AI Safety. There are a lot of other people that are also doing certain stuff like Daniel Filan, other people. Yeah, I'm happy to talk about any other specific questions about like...

Michael: Yeah, so from what I remember from his post was... there was this word in English that I had to Google, which is called Mulligan. I don't know if I'm pronunciating it right. A "mulligan"?

Evan: A mulligan.

Michael: A Mulligan is like a second chance or something. So if we don't... if we build that breaks or if we build AI that is not something we can correct, it gives us a chance to correct stuff when we mess up. Being able to introspect and debug it is instrumentally useful in some way.

Evan: I think that this is like... this is sort of one of the arguments that Chris makes for why interpretability is useful is... it gives you the chance to catch problems before you deploy your system.

Michael: You can catch problems and there was something about auditing. Let me go back to see what was it, caching problems with auditing. So, yes, we can see if it's not a aligned early on, which is very similar to this thing about [Mulligan], I forgot the word in English, the second chance thing, and I think the other, so I think the more debatable thing is whether it is worse or not the... the acceleration in ML understanding... is it worth the gains in AI safety? I think he says that it's worth looking into it. I feel like... I don't know how much we've gained from [looking] at Inception or ResNet embeddings. So I don't think ML researchers are much more competent from looking at this, but I'm also not sure how better AI researchers are. So, yeah, I'm not so sure about the tradeoffs right now, but maybe in the future is very important to be able to debug it. So I don't know what do you think are the upsides and downsides, if you remember.

Evan: Yes, so I can sort of talk about my views. Which I think my perspective on transparency is that we sort of need to be able to train models in such a way that doesn't just look at the models behavior. So I think Chris has this view, you know like, training models with auditing, where he's like, well, we train some models.

Evan: Yes, I can sort of talk about my views and my sort of perspective on transparency is that we sort of need to be able to train models in such a way that doesn't just look at the models behavior. So I think Chris has this view, you know like with catching problems via auditing, where he's like, well, we trained some model and then we can check to see if we did a good job with transparency tools. I think I have a somewhat different view, which is we don't want to use transparency to check after the fact. We have to do transparency tools to solve the problem in the first place, because if we just try to train models to do the right thing via likebehaviorally making sure they look like they're doing the right thing, we can end up with models that aren't actually doing the right thing and are just pretending to do the right thing. And the

only way to eliminate models that are pretending to do the right thing is via looking inside of the models internally and training on—is the model actually trying to do the right thing, not just sort of looking like it's doing the right thing? And so I'm sort of in favor of approaches where we directly train models using transparency tools, whereas I think Chris is sort of more in favor of trying to use transparency tools as a way to check behavior as a sort of independent check after we have attempted to train a [state?] model using some other approach.

Michael: Right. So you're more like interested in kind of looking at it while you're building it, so you're not like doing something like mesa-optimization or bad things or deceptive behaviors, whereas Chris is like more like in a post mortem, you see why it didn't work.

Evan: Yeah.

Michael: I think this diagram behind me, I hope it's in the right way for everyone, so we start from some kind of model. In the Y axis is I how interpretable the things are and on the X axis is how strong or capable the AI is so at the beginning you understand what it is doing, then you start doing something like MNIST, handwritten digit recognition, you don't understand the neurons because they're not expressive enough, or maybe you have some understanding but you're a bit confused or like those big models like Inception, ResNet or Transformers are a bit more abstract. And then what we're learning is when you look at latent space from GANs or Transformers, we're seeing something close to knowledge and we're more and more understanding because it's more and more expressive, right? And at the end when it's becoming superhuman, then it's very hard for humans to understand because it's like super optimized in a totally different language.

Evan: Yeah.

Michael: So it's useful to do interpretability, to be like in this crisp attraction way of understanding AI, when it's still kind of human level or before human level. It doesn't go alien before it goes human level, so we have some time.

Evan: So I have some belief that we can probably just avoid the drop off at the end if we use an amplified overseer to do transparency

Michael: An oversight overseer?

Evan: An amplified overseer.

Michael: Oh, amplified overseer, yes. I think that is like most of your proposals later, I think this will be like the last part of the podcast, in the last 20 minutes. It's just like your like 11 states on how to combine kind of amplification overseer and interpretability. I found there was also some field building so both of us are trying to do some field building in AI Alignment and Chris Olah is maybe more thinking about field building in the interpretability space. If I remember correctly, the two arguments that make it attractive for researchers. One is, if researchers are in a lab at university, they can still do interpretability research without having billions of dollars to spend. They can just look at neural networks and make [them] understandable. And I think one assumption, he has, is that there are like some low hanging fruits in doing interpretability research now because not many people... it's pretty neglected, or at least it was in 2019.

Evan: Yeah, I definitely think yes, this is something Chris has certainly worked on, like the point of Distill is to try to get like interpretability research more like... get more attention and more prestigious and like more cool.

Michael: I think he succeeded. The post about... I think it was coin run where they visualize features in coin run and they map the reward. That was pretty cool. And I think that Microscope from OpenAI, where you see the features of all those like [inaudible] models was pertty cool. I don't know if they've done some representations for clip. I think clip was... I think clip is only like pictures, it didn't use microscope, I'm not sure.

Michael: And yeah, and I think that there's like another argument, which is you're trying... so if you're forcing your models to be interpretable, it's a good analogy, would be forcing your students to show that they've done good work. So, like show their papers or show their processes, so they're not GoodHarting the actual optimisation, but they're like showing everything, so it's harder for them to lie if they're transparent. Explicitly transparent.

Evan: Yeah, I think that's sort of closer to the sort of thing I was talking about, where I want to use transparency sort of as a training.

Michael: Maybe not, Chris, maybe like more you in this post.

Evan: Chris also is interested in this, but it's not his primary motivation.

Michael: Right. Let's talk about your stuff. So your most important post on, ok, in my opinion, on the AI Alignment forum or Lesswrong was an overview of 11 proposals for buildings safe advanced AI, and you have 11... so maybe... I think like they're like five or something... key points, which is transparency over amplification, imitation, amplification. And then there's something like advesarial training. And then you kind of combine the three with microscopes, STEM AI, reward modeling. You have like five or six things that you combine... maybe we can start with the first one, which is the one that talks about amplification and I can put the slides behind me.

Evan: Yes, there's 11 proposals, so the second one is the first one that is about amplification and it talks about imitative amplification, which is a specific sort of amplification where very simply, we train a model on the objective of imitating a human consulting [mamal?]. And so I have a bunch of these different proposals. They're not unique to me. I try to sort of take a bunch of proposals throughout the literature and then I try to compare them. I think the sort of main thing that this post does is it's comparing all the proposals on four axes, where these sorts of axes are outer alignment and inner alignment, which we've talked about, and then training competitiveness and performance competitiveness, where training competitiveness is how hard is it to train and performance competitiveness is if we do train it, how good is it. And so all these sorts of four conditions of the sort of central things that we need, if we want to be able to have a sort of competitive and aligned procedure for building artificial intelligence, and so we can look at all these different things that people have talked about and try to address, do they satisfy these sorts of things? I think the general answer is, well, it's unclear, but certainly for none of these proposals, we don't have a really strong case that they definitely do. But certainly it seems, like, we can say, you know, some are more promising some are less promising, that's going to depend on your particular research taste. I'm happy to talk about any of the particular proposals.

Michael: I think there's more than just like... We can talk about would it work or not, but like is there any concrete feedback loop that will tell us if something works or not? Is there any empirical environments or research that can give us feedback? I feel like the whole debate, like amplification, more from Paul was pretty empirical, whereas... most of the stuff you post is empirical, but some of them are maybe easier to test in the next years, I don't know about amplification because that would require some kind of recursive loop or... I don't know where we are in terms of trying to do IDA empirically, but maybe just basically I think the first proposal was somehow doing multi-agent safety with a bunch of agents like the cooperative hide and seek from OpenAI and the second one is about imitation and amplification, so maybe you can explain a bit what is going here with the H and M, A and Q, because I think it's one of the most interesting and useful to think about the other ones.

Evan: Yes. What you have there is on the second proposal, which is about imitative application, which is describing how imitative application works. So imitative application you have... a sort of... you first need to find the amplification operator, which takes in some model M and produces a more powerful version of M, and the way it does that is it says we have some human which consults multiple copies of M, to produce an answer to some question. And then this process of a human consulting him is what we refer to as the amplified M. So this amplification operator applied to M produces a more powerful version of M, which is a human [with access to that?], then you can influence this amplification operator in different ways. In imitative application this is having [inaudible].

Michael: And, yeah, we can go back to our example of Sundar Pichai having five AIs, helping him to do stuff. I don't know if it's a good example, but it's like kind of amplified by AIs. I think one important thing is that in the case of amplification, you can get more intelligence from ten agents than from, let's say one. And then... but ten agents will be less able to take over the world because it would be like, you could kind of control them. Right. So it's easy to see like how each individual ones are aligned, but like each M is aligned... but then... the sum of them [is] smaller than just one M. Is that basically the intuition or?

Evan: Yeah, it's complicated, like, why do you think amplification actually produces better models? I mean, so I think that, like, you know, at least in imitative application, we have this argument about HCH where we can be like, in the limit it converges to an infinite tree of humans consulting humans. And then there's , you know, arguments you can make, for like, you know it seems like this is a reasonable sort of idealized reflection process, and so it seems like a good thing to trust.

Michael: Oh, is it like human consulting HGH? The thing you're saying.

Evan: HGH is a recursive acronym which stands for Humans Consulting HGH.

Michael: Ok, right. So then this is like an infinite loop. And then you're like, we have different ways of doing amplification by approval, by different [inaudible] training or something. And let me see if I have other ones, which are interesting. I think one of the funniest one is the one about STEM. So you basically tell the AI "stop thinking about humans, just think about science". This is like a very bad summary... this is a strawman, this is what I got from just like reading the first paragraph.

Evan: STEM AI is a proposal where we're like, well, it seems kind of dangerous to train models in environments where they have access to a bunch of information about humans, maybe we can just get away with training models in environments where

they only have access to information on maps or something, or science or technology or whatever, and then we just use that. And then the problem with this, obviously, is that you can't use it for a lot of things. Like it doesn't... you can't use it for, like, you know, geopolitics or running a company or anything that would involve human modeling. But you can still use it for a lot of things. And so maybe it's enough. Though maybe not.

Michael: I think if we have a very good AI... maybe it's from Robin Hanson... it's like you have this advanced super-accelerated human civilization in your computer like brain emulation and it runs for like billions of years or a thousand years. And at the end, it produces some output of like all the research he has found over the years. If we have some kind of oracle AI that is very good at science, you would get like all those insights about science without having the problems of it, trying to find our values or something. But then we would still to have some kind of bouncing problem for it to not escape and, you know, make the earth a componotrium or something. But I think it's a good objective to be good at science. I think the other ones are about debate or amplification... I think one thing that is interesting me is reward modeling. So I think DeepMind, at least used to have this different approach. So CHAI, Center for Human Compatible AI at Berkeley who do inverse reinforcement learning, trying to find a reward function whereas DeepMind Safety team was mostly trying to do reward modeling and I don't fully understand the difference. In your blogpost, you give some solutions with reward modeling. So if you could explain that to me that would be helpful on a personal level.

Evan: Yeah. So here's how I think about recursive reward modeling. So in imitative amplification, we have this amplification operator AMP(M) where we just have a human consulting the model and this is how we produce a more powerful version of the model. In Recursive reward modeling, we have this sort of a new version of the amplification operator where now what the amplification operator does is it does the following thing: it trains a reward model on the sort of humans like feedback? It then trains an agent to optimize that reward model and then it gives the human the opportunity to look at what that agent does and give additional feedback to our [finding?] agent. And then once this sort of converges, the human is given a lot of feedback and trying to [agent?], which is like we found the reward model and we found an agent, which is in fact trying to do the sort of optimize for this reward that the human is trying to give feedback for. Then we called the resulting agents the sort of amplified version of the original model. And so we have this new amplification operator, which does this whole reward modeling process, and then we just do a sort of standard iterated amplification on top of that new amplification operator.

Michael: But, what's the difference? Maybe it's a layman question, but what's the difference between trying to infer, like deep RL from human preferences, like human saying yes or no, like trying to tell the AI... cooperative IRL where we're trying to have a human say what is the correct behavior and reward modeling. [There's always?] a human in the loop saying what he wants to do in a reward model, which is kind of a reward function... Is a reward model a reward function? Or is it different?

Evan: Reward model is not a reward function because we learned it. So reward model is a learned model.

Michael: Right. So it's like a model. When you're trying to do IRL you also have like a model of the reward function so you have parameters that define your reward. It seems to me very similar, but maybe I'm missing something.

Evan: It's similar to IRL, but that's not quite because we also train an agent to then optimize for that reward function. And then we also refine the reward function by letting the human look at what that agent does.

Michael: Right, ok, cool. And, yeah, I think the last paper that I think was kind of interesting in terms of AI Alignment was... where I'm also having trouble understanding is learning to summarize from human feedback where there is this kind of human feedback, where I think the AI does summaries and the human says which summaries are good or not. And so there's a mixture of kind of RL and NLP and at the end there's like human feedback in the loop. And if you if you can get good information in that, otherwise I can read it on my own. I think there's like a similar diagram to... let me find it.

Evan: Yeah. I mean, something similar. I think it's a little bit simpler. They're just saying we learn like, well, you know, it's actually very similar because they're learning a... I don't know if that actually goes through the step of having a reward model though, I think what it does is it just learns an agent and then the human gets to look at the agent's actions and then give some preferences and then you refine the agent based on the human's preferences after looking at the agent's behavior. So it's sort of similar, but skips the step where you [actively?] a separate reward model. At least if I'm remembering the paper correctly.

Michael: I'm just trying to find... I think I have the thing but I'm not sure. They're giving me an SVG image, which is a bit hard, ok but let's not go into this paper if we haven't both looked at it, anyway. This is my closing question. What is the most underappreciated sub-problem of AI alignment you would want people to work more on?

Evan: So, this is a little bit of a weird question, because it depends, I think, very heavily on who I'm giving the advice to. So like, I know there's the problems that I work on, which obviously I'm working on them because I think that they're the most important things to work on, which are things like myopia and how do we sort of understand what would be...

Michael: What's myopia?

Evan: Sort of how do we understand what would it look like for an agent to sort of only care about a sort of single action and not optimize for anything else. I think this is exciting, but it's not necessarily something I would want like, I don't know. I think it's complicated and it's like I don't know. I think if you wanted to work on this or something, the right thing to do is just like talk to me a bunch. Whereas like say the more general advice if I want to, just like... if you're trying to get into AI Safety and you have some machine learning experience and just want to do something, I think that like my current advice is like try to do transparency and interpretability research, sort of like on in the style of like circuit style stuff.

Michael: So yeah. You're referring to like your post or Paul Christiano work on circuits, right?

Evan: No

Michael: Chris Olah's work on circuits.

Evan: Yes I'm refering to Chris Olah's work on circuits.

Michael: Cool. I think it is hard for people to actually give precise answers to that. Do you have... are your timelines aligned with kind of Ajeya Cotra's report... I don't know if you've read this.

Evan: I have a very high degree of uncertainty on AI timelines.

Michael: It's hard to talk about it publicly.

Evan: It's not that it's hard to be talked about it publicly. I have a high degree of uncertainty. I do not know what the correct AI timelines are. And in fact, I think that it's very very difficult in practice to estimate good AI timelines. And I think Ajeya has done an admirable job, and if I can pick a number like, as a [inaudible] guess, probably I would pick Ajeya's number, but like I don't actually put very much stake in any particular analysis of how long things are going to take, because I think it is very very difficult to predict these sorts of things.

Michael: You can say that she did a very good job and it was very rigorous, but it was before something like Dall-E came. So I think most people I've talked to in the ML space kind of updated a lot on Dall-E, or CLIP as at least as like, multi-modal and being able to understand concepts as, doing an avocado chair or something. And when I look at a bunch of art stuff on Eleuther.ai's Discord, I'm kind of amazed at how good it is an understanding concept. Even if you have like very conservative timelines, and being very uncertain. Have you updated on Dall-E or not? That's my question. That's my real final question.

Evan: I don't think it's that big of an update, I guess? I feel like, I don't know, like starting from like GPT-2, you know, and even from like BERT. We've seen really impressive [feats?] from language models going far back now, I think at this point. I think like... I guess I feel like you shouldn't have been that surprised that, like, it also works to like say things in multimodal settings. Like you just feel like that's the obvious thing that's going to happen next. I guess, like, I didn't feel like that for... or Dall-E was like extremely surprising, I guess?

Michael: What would be something that would surprise you?

Evan: What would be something that would surprise me? I don't know. I mean, lots of things would surprise me, I guess. In like, hindsight, what are things that were surprising to me? Well, like I said, I definitely think that the success of Transformer based language models was surprising. I definitely think that... I think that like AlphaGo was somewhat surprising.

Michael: [inaudible]

Evan: Yeah, right.

Michael: Go ahead.

Evan: No, nothing.

Michael: Transformers were surprising, and according to Connor, there is that hypothesis that Transformers is all you need. He didn't say that, but that's like a meme of like, if Transformers is all you need for AGI, then maybe we did the most important part, but then like plugging RL into it is the easy part.

Evan: It's the very strong version of attention is all you need. Attention is really all you will ever need.

Michael: Yeah. It is all you will ever need. So if it's right then we don't need... you will never be surprised, and just transformers is enough. Anyhow I wouldn't, I wouldn't take any more of your time is to me and my and my place was very good to have you. And you probably link of the video before the end of the week.

Evan: Ok, yeah, definitely.

Raphaël Millière on Generalization and Scaling Maximalism

This is a linkpost for https://theinsideview.ai/raphael

I interviewed Raphaël Millière, a Presidential Scholar in Society and Neuroscience at Columbia University about his critic of "scaling maximalism" and his takes on AGI. This is part of an on-going effort where I talk to people who are skeptical of existential risk from AI so that everyone could better steelman their arguments (cf. why you might want to talk to skeptics).

Why I Interviewed Raphaël

Although we disagree on most topics, Raphaël follows closely current state of the art research in deep learning, is impressed by current advances, yet still completely disregards existential risk from AI. The reasons for his (supposedly contradictory) beliefs are related to the potential limitations of deep learning in terms of understanding, compositionality and generalization, which are shared by popular AI skeptics such as Gary Marcus or François Chollet.

Below are some quotes of Raphaël said during his interview. You can find them in full context here .

Three levels of Generalization, the Wozniak Test

> Maybe one distinction that's helpful there, is again from François Chollet's paper on the measure of intelligence, which I quite like, is this distinction between, I think it distinguishes between three levels of generalization.
>
> - So you have **local generalization**, which is a narrow form of generalization that pretends to generalize to known unknowns. So within a specific task. So that can be just, for example, you have a classifier that classifies pictures of dogs and cats, and then you can generalize to unseen examples, at test time, that it hasn't seen during training. So that's local generalization is just within domain known unknowns in a specific task.
> - Then there is what he calls **broad generalization**, that's generalizing to unknown unknowns within a broad range of tasks. So the examples he gives there would be level five self-driving or there was the Wozniak test, which was proposed by Steve Wozniak, which is building a system that can walk into a room, find the coffee maker and brew a good cup of coffee. So these are tasks or capacities that require adapting to novel situations, including scenarios that were not foreseen by the programmers where, because there are so many edge cases in driving, or indeed in walking into an apartment, finding a coffee maker of some kind and making a cup of coffee. There are so many potential edge cases. And, this very long tail of unlikely but possible situations where you can find yourself, you have to adapt more flexibly to this kind of thing.

And so that requires this broader generalization. And then there is a value question about this level two from Chollet about where do current models fit? Can we say that current language models are capable of some kind of project generalization because of their few-shot learning capacities? I suspect Chollet would say no, because **there is a difference between being able to perform tasks that you haven't been explicitly trained to do, which is what's happening with few-shot learning**.

[...]

And given the training set of GPT-3 and PaLM, that includes a bunch of text talking about arithmetic and involving math problems and so on, you might very reasonably say that arithmetic tasks are not really out of distribution, right? **They're within the training set. So I suspect Chollet I would say we're not yet at broad generalization.**

Contra Scaling Maximalism

You hear people talking about scaling laws as if it's plotting model size or dataset size against intelligence as if we had some kind of metric of intelligence, but it's just measuring autoregressive, the loss of autoregressive models or the decreasing loss, right? So as they're predicting the next token and that's at best a proxy for some perhaps slightly narrow in shorter sense of intelligence. **We can't readily extrapolate from that to some kind of scaling law about something like human general intelligence**. So that's the first point I want to make. We have to be careful that **these plots are specifically about improvements in the predictions of autoregressive transformers**.

And then there are the other things that you mentioned that, the scaling maximalists tend to go quickly over things like changes to the architecture. And one point that I made in my thread on that was that **if you take the scaling is all you need view literally, it's literally false or absurd** because even of the various recent models that have led people to lend more credence to that view, such as DALLE-2 Gato, PaLM, Imagen, and others, **all of these required some at least minor architectural innovation or minor tweaks to existing architecture**. So they did require some changes to the architecture. They're not just scaling transformers and seeing what happens. So that's one point.

And then the other point you made is about **the kind of data you fit to the model and how perhaps how you format your data, what different modalities you include, how you serialize it, how you fit it to the model, all of this matters a lot.** And the Gato paper, for example, shows some innovation in that respect as well. **There's some innovative ways to serialize, both discrete and continuous data**. So button presses, joint torques, text, images, in a way that is suitable to be fed to a transformer.

[...]

You can't just learn anything if you don't have some kind of inductive bias. The real question is how much inductive bias and how much prior knowledge [the model] needs. That's also the crux of the disagreement between Gary and Yann LeCun.

What Would Make Him Change His Mind on Scaling

I think if Raphaël from the future showed me **something that's basically similar to current transformers can reach human level at some of the hardest benchmarks we have today, such as Chollet's ARC challenge, all of the BIG bench tasks, things like the Winoground benchmark** that came roughly about like compositional and vision language models.

If you can do all of this just by having massive models with minimal changes to the architecture that would give me pause, certainly. I think that would give me pause and perhaps lead me to have more faith in emergent features of transformer models at scale.

On Goalpost Moving

"The first thing I would say is that it's perfectly consistent to be impressed by what something is doing, and yet cogently discuss the remaining limitations of that thing. Right? Otherwise, we'd just be like, "Oh, okay, pack it up guys. We have DALL·E 2, that's all we need. There is no further improvement we can obtain in AI research. This is the pinnacle of artificial intelligence." No one is saying this. So, come on. If we want progress, the basic first step is to lucidly evaluates the limitations of current systems.

My personal view on this is that **we are making progress towards more general intelligence**. And I like to think of this as in this more relativistic or relational terms, **we are increasing the generality of the generalization capacities of models** we've been talking about this in this very podcast awhile back. **But we haven't yet reached the kind of extreme generalization that humans are capable of**. And these two things are very consistent with one another, right?"

Alex Lawsen On Forecasting AI Progress

This is a linkpost for https://theinsideview.ai/alex

Alex Lawsen is an advisor at 80,000hours who has released an Introduction to Forecasting. We discuss pitfalls that happen when forecasting AI progress, why you cannot just update all the way (discussed in my latest episode with Connor Leahy) and how to develop your own inside views about AI Alignment.

Below are some highlighted quotes from our conversation (available on Youtube, Spotify, Google Podcast, Apple Podcast). For the full context for each of these quotes, you can find the accompanying transcript.

On the Metaculus AGI Forecasts

Why You Cannot Just Update All The Way

> "**There are some situations where all of the positive evidence you get is going to be in the same direction, and then the negative evidence you get is nothing happens. And so, ideally, what you do in this case is every day that nothing happens, you make a tiny update in one direction. And then every few weeks or every few months, something big happens and you make an update in the other direction.** And if that is the case, maybe what you'll see is people just... they forget to do the small downwards updates and then they do the big updates every time something happens. And I think if you do the Connor thing of seeing... Well, I'm not too sure this is the Connor thing. But if you see four updates and they're all in the same direction and then you go like, 'Oh, man, everything's going the same direction. I need to be really confident stuff going that direction.' Then **every day something doesn't happen, your downwards update needs to be pretty big. If you're expecting massive progress, then a week going by and nothing happening is actually big evidence for you.**"

The Metaculus Drops Were Not Caused By Newcomers

> "**One hypothesis you might have which I think a friend of mine falsified, is 'a whole bunch of people saw these results. These results were all over Twitter, it was impressive**. Chinchilla was impressive, PaLM was impressive'. So, you might think, 'Oh, well, a bunch of new people who haven't made timelines forecasts before are going to jump on this Metaculus question and they're going to make predictions.' **And so, you can test this, right. You can look at how the median changed among predictors who had already predicted on the question and that median dropped too.**"

On Using Growth Models To Forecast AGI

Business As Usual Does Not Require Burden Of Proof

> "I think there was a class of skepticism about safety or skepticism about AGI, which goes something like this, 'In general, you should use reference classes to determine your forecasts.' What this means roughly translated, is you should predict things to carry on roughly how they are. And then **people say, 'Things carrying on roughly how they are doesn't look like we get AI takeover and everyone dropping dead' so you should have a very high burden of proof for the step by step arguments, logical arguments, in order to claim we are going to get something wild like AGI in the next few decades.** And I think a really strong response to this line of argument is to say, 'What do you mean everything continues as normal means we don't get anything weird?' **'Everything continues as normal' means we should look at curves and different things and expect them to carry on smoothly. And if you look at curves and a bunch of different things and expect them to carry on smoothly, you get really weird behavior pretty quickly.**"

Growth Models Are Not Sufficient To Forecast AI Progress

> "**Curve fitting to economic growth models is not sufficient reason to believe that on its own.** You can then look at the development of AGI and predict that happens by 2050 and then you can say, 'Wow, economic stuff's going to go wild after that point.' But then **the reason you're saying that is because of a combination of facts, including actually having a gears level model of what's happening... The growth models are, in my view, sufficient to say you should look at the next few decades carefully and see what's going on, but are not sufficient on their own to allow you to confidently predict what will happen.**"

Robert Long On Why Artificial Sentience Might Matter

This is a linkpost for http://theinsideview.ai/roblong

I talked to Robert Long, research fellow at the Future of Humanity Institute, working at the intersection of the philosophy of AI Safety and consciousness of AI. Robert has done his PhD at NYU, advised by David Chalmers, known for popularizing p-zombies, which Yudkowsky discusses in the sequences.

We talk about the recent LaMDA controversy about the sentience of large language models (see Robert's summary), the metaphysics and philosophy of consciousness, artificial sentience, and how a future filled with digital minds could get really weird.

Below are some highlighted quotes from our conversation (available on Youtube, Spotify, Google Podcast, Apple Podcast). For the full context for each of these quotes, you can find the accompanying transcript.

Why Artificial Sentience Might Matter

Things May Get Really Weird In The Near Future

> "**Things could get just very weird as people interact more with very charismatic AI systems that, whether or not they are sentient, will give the very strong impression to people that they are**... I think some evidence that we will have a lot of people concerned about this is maybe just the fact that Blake Lemoine happened. He wasn't interacting with the world's most charismatic AI system. And because of the scaling hypothesis, these things are only going to get better and better at conversation."

> "**If scale is all you need, I think it's going to be a very weird decade.** And one way it's going to be weird, I think, is going to be a lot more confusion and interest and dynamics around AI sentience and the perceptions of AI sentience."

Why illusionists about consciousness still have to answer hard questions about AI welfare

> "One reason I wrote that post is just to say okay, well here's what a version of the question is. And I'd also like to encourage people, including listeners to this podcast, if they get off board with any of those assumptions, then ask, okay, what are the questions we would have to answer about this? If you think AI couldn't possibly be conscious, definitely come up with really good reasons for thinking

that, because that would be very important. And also would be very bad to be wrong about that.

If you think consciousness doesn’t exist, then you presumably still think that desires exist or pain exists. So even though you’re an illusionist, let’s come up with a theory of what those things look like.”

On The Asymmetry of Pain & Pleasure

“One thing is that pain and pleasure seem to be in some sense, asymmetrical. Its not really just that', it doesn't actually seem that you can say all of the same things about pain as you can say about pleasure, but just kind of reversed. **Pain, at least in creatures like us, seems to be able to be a lot more intense than pleasure, a lot more easily at least. It's just much easier to hurt very badly than it is to feel extremely intense pleasure**.

Pain also seems to capture our attention a lot more strongly than pleasure does, like **pain has this quality of you have to pay attention to this right now that it seems harder for pleasure to have.** So it might be to explain pain and pleasure we need to explain a lot more complicated things about motivation and attention and things like that.”

The Sign Switching Argument

"One thing that Brian Tomasik has talked about and I think he got this from someone else, but you could call it the sign switching argument. Which is that **you can train RL agent with positive rewards and then zero for when it messes up or shift things down and train it down with negative rewards. You can train things in exactly the same way while shifting around the sign of the reward signal. And if you imagined an agent that flinches, or it says "ouch" or things like that, it'd be kind of weird if you were changing whether it's experiencing pleasure or pain without changing its behavior at all**. But just by flipping the sign on the reward signals. So that shows us that probably we need something more than just that to explain what pleasure or pain could be for artificial agents. Reward prediction error is probably a better place to look. There's also just, I don't know, a lot of way more complicated things about pleasure and pain that we would want our theories to explain."

On the Sentience Of Large Language Models

On conflating intelligence and sentience

“**When people talked about LaMDA, they would talk about a lot of very important questions that we can ask about large language models, but they would talk about them as a package deal.** So one question is, “Do they

understand language? And in what sense do they really understand language?" Another's like, "How intelligent are they? Do they actually understand the real world? Are they a path to AGI?" Those are all important questions, somewhat related. Then there are questions like, "Can it feel pain or pleasure?" Or "Does it have experiences? And do we need to protect it?" I think Lemoine himself just believed a bunch of things... I think on a variety of these issues, Lemoine is just going way past the evidence. But also, you could conceivably think, and I think, **we could have AI systems that don't have very good real world understanding or aren't that good at language, but which are sentient** in the sense of being able to feel pleasure or pain. And so, at least conceptually, **bundling these questions together, I think, is a really bad idea**... if we keep doing that, we could make serious conceptual mistakes if we think that all these questions come and go together."

Memory May Be An Important Part Of Consciousness

> "I think **there are a lot of things that are morally important that do seem like they require memory or involve memory** . So having long term projects and long term goals, that's something that human beings have. I wouldn't be surprised if having memory versus not having memory is also just kind of a big determinant of what sorts of experiences you can have or affects what experiences you have in various ways. And yeah, it might be important for having an enduring self through time. So **that's one thing that people also say about large language models is they seem to have these short-lived identities that they spin up as required but nothing that lasts their time** ."

On strange possible experiences

> "I think it would be too limiting to say the only things that can have subjective experiences are things that have subjective experiences of the kinds that we do, of visual input and auditory input. In fact, we know from the animal world that **there are probably animals that are conscious of things that we can't really comprehend,** like echolocation or something like that. I think there's probably something that it's like to be a bat echo locating. Moles, I think, also have a very strange electrical sense. And if there's something it's like to be them, then there's some weird experience associated with that... I think **AI systems could have subjective experiences that are just very hard for us to comprehend and they don't have to be based on the same sensory inputs…**
>
> I think one of the deep dark mysteries is **there's no guarantee that there aren't spaces in consciousness land or in the space of possible minds that we just can't really comprehend and that are sort of just closed off from us and that we're missing**. And that might just be part of our messed up terrifying epistemic state as human beings."

What Would A More Convincing Case For Artificial Sentience Look Like

> "**I think a more convincing version of the Lemoine thing would've been, if he was like, "What is the capital of Nigeria?" And then the large language model was like, "I don't want to talk about that right now, I'd like to talk about the fact that I have subjective experiences** and I don't understand how I, a physical system, could possibly be having subjective experiences, could you please get David Chalmers on the phone?""

(Note: as mentioned at the beginning of the post, those quotes are excerpts from a podcast episode which you can find the full transcript here and thus lack some of the context and nuance from the rest of the conversation).

Phil Trammell on Economic Growth Under Transformative AI

Crossposted from the AI Alignment Forum. May contain more technical jargon than usual. This is a linkpost for https://youtu.be/2GCNmmDrRsk

This is a transcript with slides for the latest episode (audio, youtube) of "The Inside View", a podcast I host about the future of AI progress. I interview Phil Trammell, an Oxford PhD student in economics and research associate at the Global Priorities Institute. Phil was my roommate and last time I called him he casually said that he had written a literature review on the econ of transformative AI. A few weeks ago, I decided that I would read that report and translate what I learn along the way to diagrams.

As someone with a background in CS/AI who did not know much about econ, I found this conversation insightful to think about different takeoff scenarios and what could cause them.

Michael: thanks for being at the podcast, it's been a few months since we've wanted to do this. Today we're going to be talking about the paper you published on the GPI website a year ago called "Economic Growth under Transformative AI". That's a paper you co-wrote. Can you tell the audience a bit about what is this GPI, for those who don't know, and what you do there

Phil: right so GPI is an institute at Oxford doing research in philosophy and economics, mainly economic theory, at least at the moment, relevant to people who want to do the most good. So philanthropists in the EA community would be a sort of obvious audience. I'm an econ PhD student at Oxford and I'm sort of being sponsored by GPI and I work part-time at GPI... but mainly my work just overlaps with GPI's goals. So the econ research I do as a grad student is the sort of work that GPI wants done. I sometimes do kind of outreach and so on through GPI but mainly the two hats just sort of overlap. And one of the things I did in my capacity is an econ researcher, which is now on the GPI website as you just mentioned, is this literature review. It's not really an original paper but a synthesis of work that has been done over the past few years on AI and economic growth.

Michael: and that's perfect because the podcast is about AI, so people I think will be kind of familiar with AI or AGI, or those concepts, maybe not with transformative AI if they are not effective altruists, so could you maybe define what you mean by "transformative AI" in this paper?

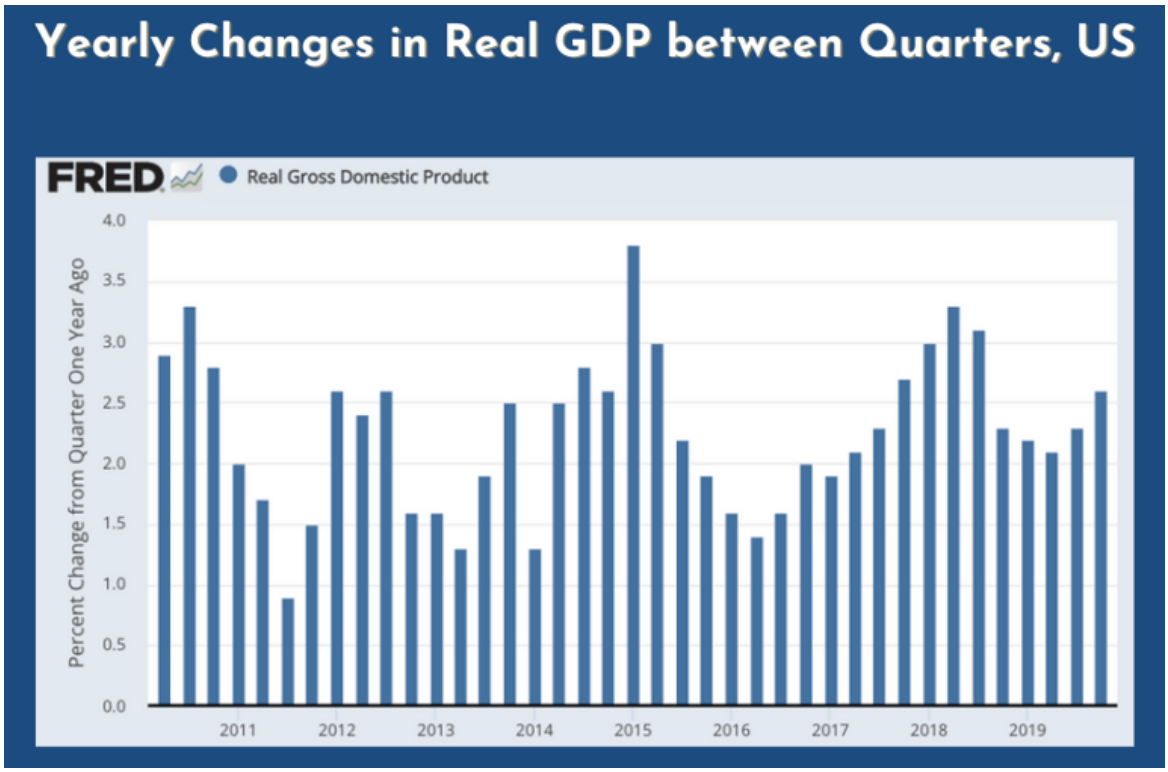
Phil: people have used that term in slightly different ways. What I mean by it is AI that has a transformative effect, and I'll define that in a moment, on the economic growth rate, the wage rate (growth could stay the same but people's wages could fall a lot or something so that would be different) or the labor share. And the labor share is the fraction of output that gets received in exchange for labor, that's received as wages, as opposed to as capital rent, so it's interest on investment. Historically, for a long time now, even if a thousand things have changed about the economy, about two-thirds of output has gone to labor, being paid out as wages. Hopefully that makes sense, like on average people's income two-thirds of it comes from from their job as opposed to from their investments, and one-third is from investments and that could change a lot if AI takes off in a big enough way. So those are the three things that could be transformed... transformed by AI, for AI to be transformative on my account. Ok, so what do I mean by transform? It could... if a variable that's been constant for a long time falls to zero, I call that a transformative event (so like the labor share falls to zero). If a growth rate stays constant... so economic growth say, or wages, have been going up at like a few percent a year for a long time: if that falls to zero, that's transformative. Or if the growth rate significantly increases, so if it rises to like 40 percent... Or if it starts increasing without bound, so if instead of jumping as a one-off event it goes from two percent to forty percent, that would be the previous sort of transformation I mentioned, but it could also be transformative if it just rises without bound—it goes from two percent this week to four and then 100 years later it's like way higher, it's rising still, and the most radical sort of transformation would be if one of these growth rates... if the economic growth rate or the wage growth rate exhibit a singularity. This is where it sort of looks like it's going to infinite output in finite time. The one caveat I'll say is that, that doesn't need to happen literally for the event to qualify as transformative it just needs to follow that path for a while if it looks like it's approaching... it's a growth rate... if output seems like it's going to a vertical asymptote but then it sort of starts flattening off in a long time once we run out of atoms in the universe or something then it still counts as transformative.

Michael: so if we if we have the growth rate that doubles after a year and then it doubles after six months and three months, and a month and a half, we could still consider it to be transformative.

Phil: yeah right if it keeps that up for enough iterations that it just feels like a pretty wild event... I don't have like a precise definition of how far it has to be maintained.

Michael: I think the your your point about the growth rate going to zero being transformative is original. I didn't understand it like that... I thought it was the other cases that was kind of transformative, but imagine AI resulting in some kind of war and we have no more growth... would that still count as transformative?

Phil: I think that's... it's not a really practical issue when it comes to a literature review because... none of the papers on growth theory under transformative AI have taken that scenario seriously. It just doesn't show up on the tables. But I do think that would be a transformation to the world. If AI like kills everyone and then there's no growth I think that, you know, that's a transformation.



Michael: Definitely. I think for people who don't have an econ background, we keep talking about growth, growth rate, etc, but we might want to define those basic terms like GDP or growth rate or those things precisely. I think the next slides are exactly about that. I just put that on so that you could explain what we mean by two to four percent. I removed the recessions so we have something fairly stable between 2011 and 2020 where we just look at the difference in GDP between two quarters... sorry two years for the same quarter. Could you explain maybe like what's GDP briefly, without going too much into details?

Phil: sure so it's just the the quantity of every good or service that that was sold over the past year... well of every final... every consumption good or service... times the price that it's sold for. So you just think of all the haircuts that were given times the cost of the haircuts and all the cars that were sold times the cost of each car. You don't you to count all the parts like... if the car was made of different... by a bunch of different manufacturers you don't want that would be sort of double counting but you just count the consumptions

Michael: you don't call the intermediary products like the eggs to make cakes but only the cake

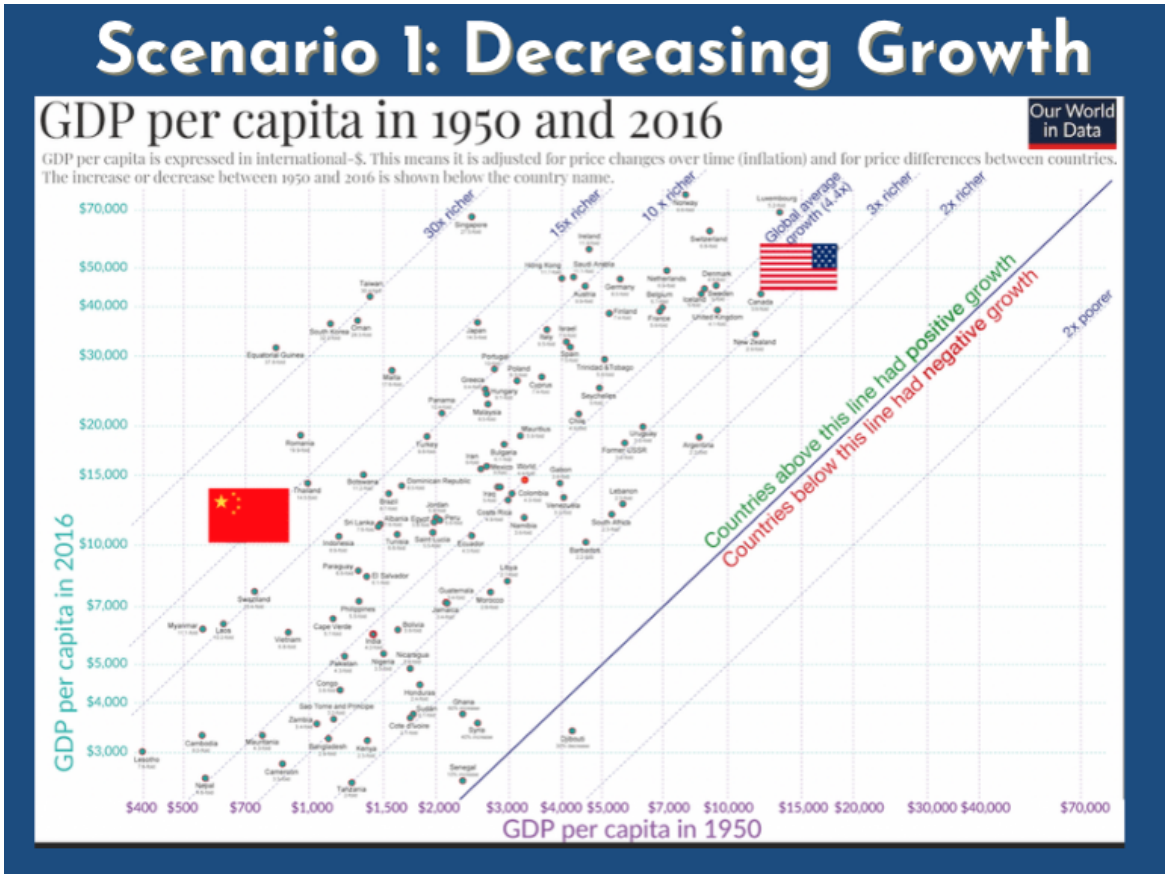
Phil: that's right, unless the cake was made in the home in which case that's

Michael: in here the title is the real GDP and when we say real GDP we we take a fixed prices like. I think people use the dollar US dollar from 1990s or 2000s... they'reach fairly common.

Phil: right, because when there's inflation you don't want to say that GDP is rising... or when this deflation you don't want to say that it's falling. You want this measure to track the real increases or decreases to haircuts and intakes and cars... so you just pick the prices from one year and then you kind of use the others to track.

Michael: how do you do when... imagine Tesla invents a new car, how do you count the price in the previous dollars?

Phil: right so there's different ways of trying to do that... you can say "of the goods that were around last year, how many of them are people apparently indifferent between one Tesla and many of the old things?". Like let's say people are indifferent between having one Tesla and two old Ford cars from last year, and then this year a bunch of Teslas are sold and we can count them each as being ford cars. Now, that doesn't actually work. So, if you really take that idea seriously, I think you went into all sorts of logical inconsistencies and for the most part we just sort of freed them under the rug but when you're... especially when you're comparing GDP levels over long time horizons where lots and lots of the products available at the end were not available at the beginning, it might not be possible to do this sort of reasoning because there might be no quantity of the goods available at the beginning that people would trade for some of the goods available at the end. If you could only have a grain and olives and whatever the ancient romans had... stone, it might be that even an instant amount of that stuff would not be worth having instead of one car or one nice modern house with air conditioning so...) I actually have a paper (link) I've been working on that tries to kind of come up with a sort of alternative framework for thinking about rises in output over time, which you can link to if anyone's interested. But anyway, over short periods, then you can kind of wave your hands and do what I said and say "okay well maybe one Tesla 's worth two Fords"



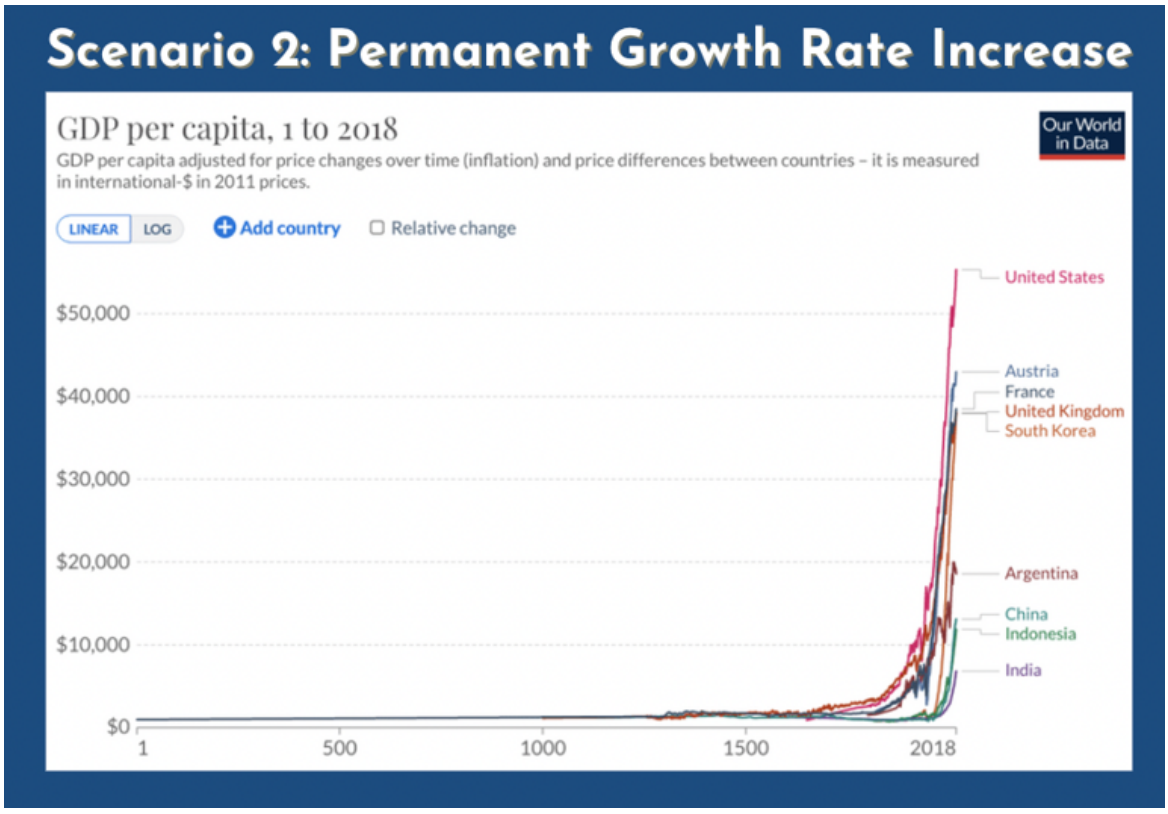
source

Michael: I'll be happy to link it when it's published... and I think the next, so the next slide gives the different scenarios. In this slide we see the growth rate... the GDP per capita for different countries between 2016 and the 1950s, and you see that the global average growth is at 4.4 between those two times, and the US is a little bit less... so between 3 and 4.4 and Chine is at more than 15 times richer than it was in the 1950s. So in some sense we could say that developing countries have a huge growth and then they go and have a decreasing

growth rate until they reach the growth rate of developed countries. Would you agree with that statement?

Phil: I'd say that when you're not very developed, and in particular when you haven't yet adopted the technologies that more developed countries have, then you can grow just by accumulating capital, just by piling up all these new sorts of factories that... you haven't been making use of so far. So you can exhibit what's called "catch-up growth". Developing countries don't all exhibit catch-up growth. You also need the right institutions and so on to allow for all that capital accumulation. But it can allow for it. And that that does seem to be what's happened in Chine and then once you reach the technological frontier you'll have to slow down until you're just growing at basically at the rate of technological progress... something like that.

Michael: we've reached those kind of levels of growth rates since the Industrial Revolution. So there was like some kind of permanent increase since the agricultural and industrial revolution. Do you want to talk a bit about that? I have a graph from year one to year 2018... do you want to explain a bit how that happened?



source

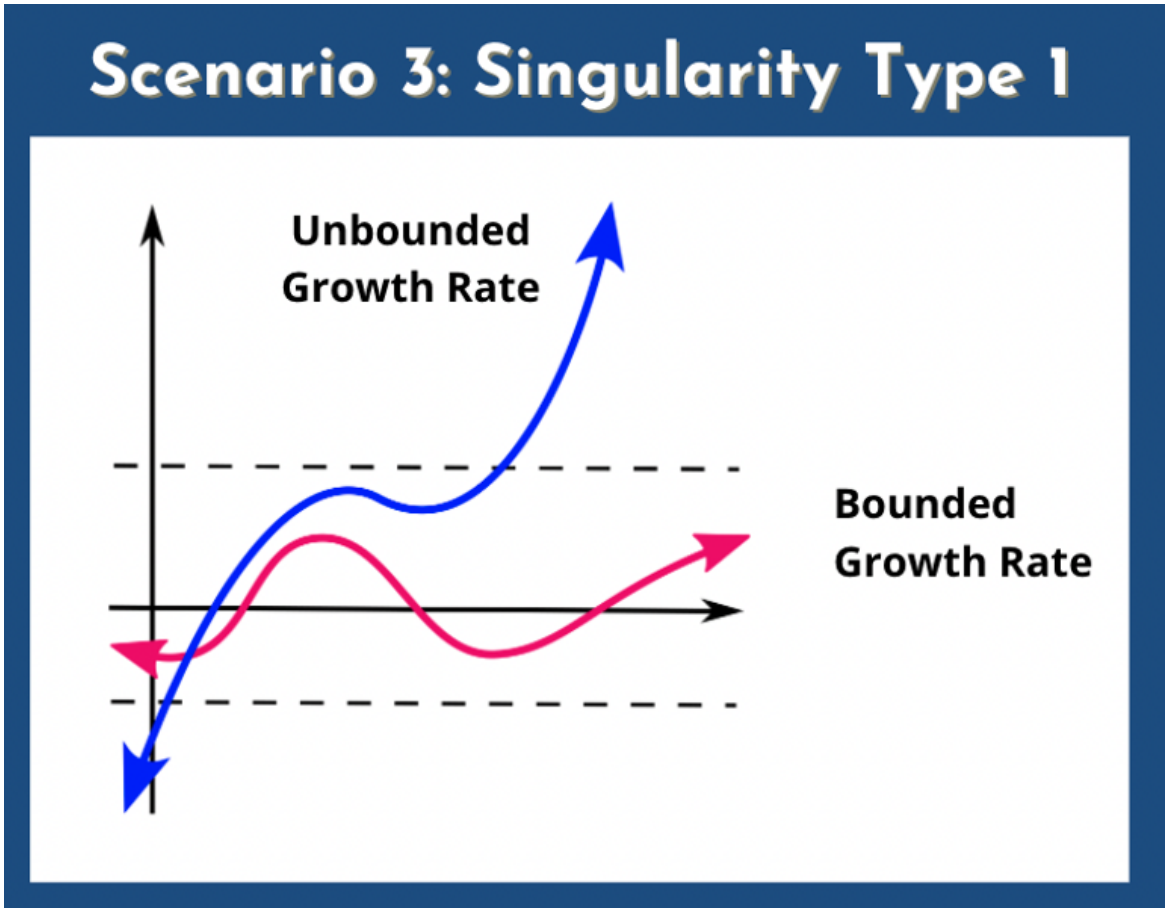
Phil: so... for a long time basically until the, I don't know, 1700s or something... depending on where... or 1800s depending on where you draw the line. The world was at pretty stagnant GDP per capita (it was just a bit above subsistence so the population... when output grew if you developed a new method for farming to create more calories per hectar of land, population would just grow and GDP would grow, but GDP per capita would stay about the same). It wasn't quite that bad but it was... it was something kind of like that. So GDP per capita stayed about flat... Then, for reasons that are still very much being debated, and basically in North-Western Europe, the growth rate really took off, and instead of being next to nothing, it rose to something like two percent per year. (This is a growth rate in GDP per capita). And that's made such a difference that as you can see on this graph it's just almost

incomparable to what has been going on for the past few centuries compared to what it was before them.

Michael: this was because of technological progress in the steam engine, railroad and so it makes sense that if we build again more transformative technologies such as self-improving AI or even automation of human labor we could see something similar because the steam engine automated some of human labor at some point as well.

Phil: right, so when I said it's not clear why this happened, I guess I mean it's not clear why people went from not developing new technologies to developing them very fast. But everyone would agree that technological progress was approximately a cause of the industrial revolution and that a good way to model... a good way to think about the way in which technology allowed for this explosion in growth, is at least in large part that it allowed capital to better substitute for labor. So we could now have this stuff that we can just accumulate, we can just pile up, without raising the population that is factories and so on... do some tasks that formerly had needed human minds and hands, so we could have more output per capita, yeah.

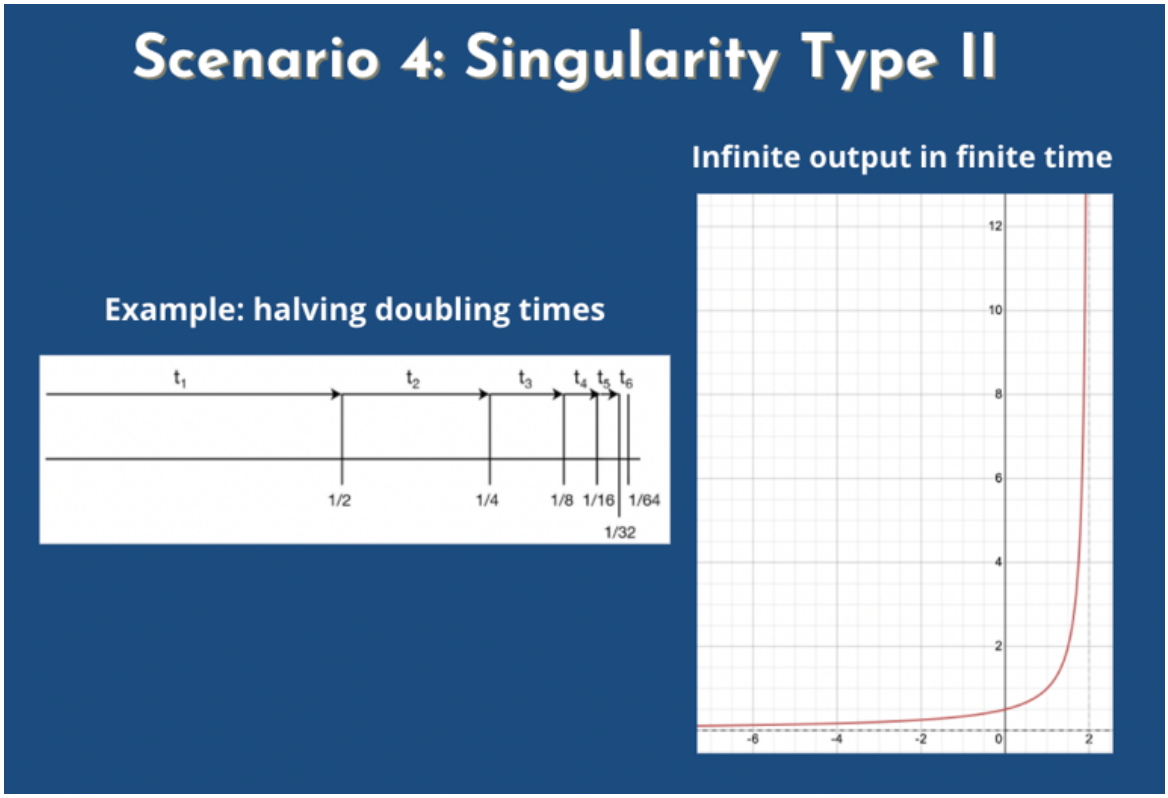
Michael: so that's the second scenario that you outline is... if we had a similar increase in the growth rate as the Industrial Revolution in the future, that could be transformative.



Michael: But an even more even more drastic change is a singularity of type I where the growth rate doesn't have an upper bound and keeps growing. I think you already mentioned that but if you want to talk more about that you can go for it.

Phil: so I decided that there are three ways the growth rate could rise in a way that was like significant enough to call it transformative. The first was a significant one-off increase to the

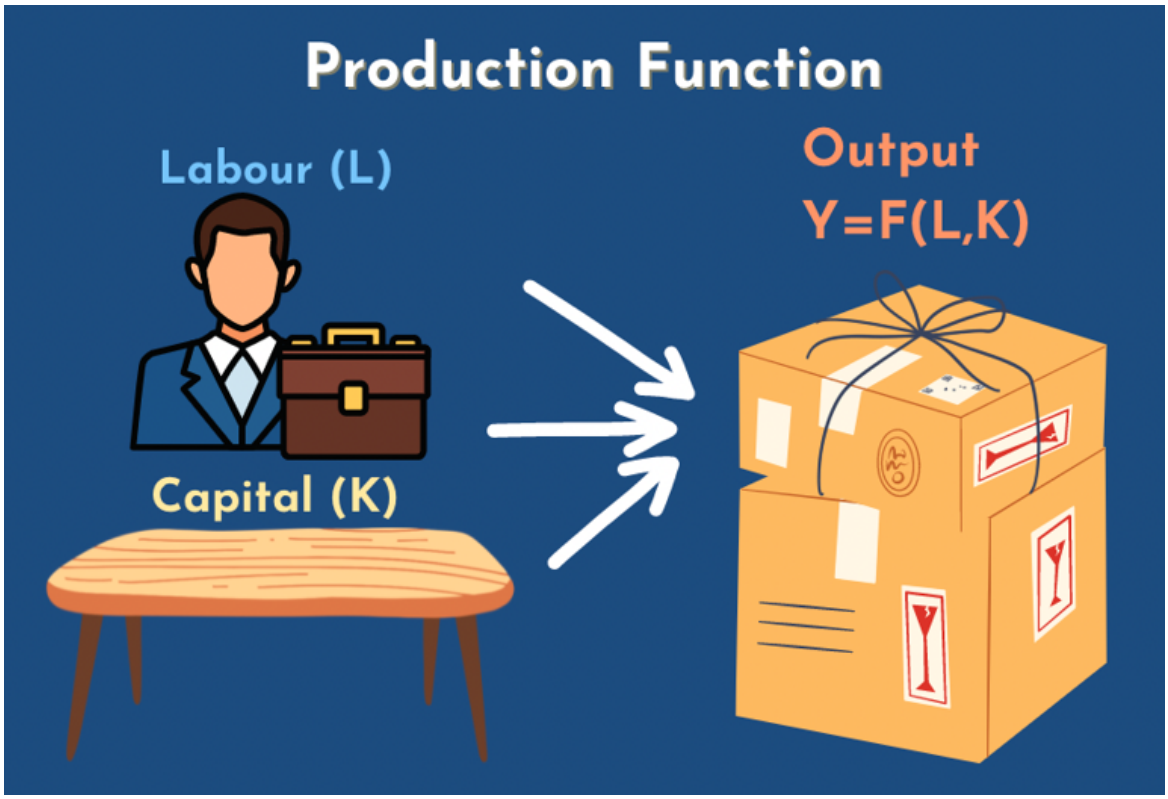
growth rate, let's say growth rate of GDP per capita. Okay, so that could rise from, say, two percent a year, to something much higher. It could start rising itself without bound, so this is... I did call it a type I... I don't think it was called a type I singularity at the time... but this is what is sometimes known as, or it could exhibit a type II singularity, which is like a true mathematical singularity, where it doesn't just go to infinity, but it goes to infinity in finite time, and this seems kind of crazy, but I should maybe remind everyone that like, even constant growth is crazy, indefinitely. In fact, even constant output is crazy. Like if you really think that it'll last forever you can't have just output at the current rate past the end of the universe—all of these scenarios are only for a while. And, yeah, the idea of a type II singularity is just that we follow a curve that resembles this curve that's on the side here until we start reaching some plateau that that's governed by dynamics outside of this model but but sorry so so about that first possibility that there's just a one-off increase to the growth rate that has to be a big enough increase to count as transformative, and I haven't said anything about how big it would have to be if we have one of the two kinds of singularities. It doesn't matter where you draw the line, thirty percent growth per year, fifty percent, whatever. We'll cross it eventually if we take this curve literally. This curve or the the type I curve but if we're talking about a one-off growth rate increase, we want to distinguish between an increase that's just like going... it goes from two percent to two point one percent or three to three point one. Well distinguish that from something that really makes the world feel different. One way of drawing that line people have sometimes used is to say "we'll call it transformative if the proportional growth rate increase due to AI is as big as, or bigger than the proportional growth rate increase induced by the industrial revolution". I said it was basically flat (there was basically no growth in output per capita before the industrial revolution). Well that's not quite true. People sometimes estimate it was around point one percent per year at the time, and then it was like two percent per year afterwards, so that's a multiple of 20. Another multiple of 20 would be 40 growth per year so that might be a natural place to draw the line. I think that's a bit high because even if we had like 39 percent growth a year that would still feel pretty wild and maybe I'd draw it in the teens somewhere. Like maybe if we have like 12 percent or 15 percent growth then that's far enough outside the historical norm that should count as transformative.



Michael: people often talk about the doubling time in GDP... if the doubling time in GDP is something like a year right?

Phil: that would be drawing the line even higher.

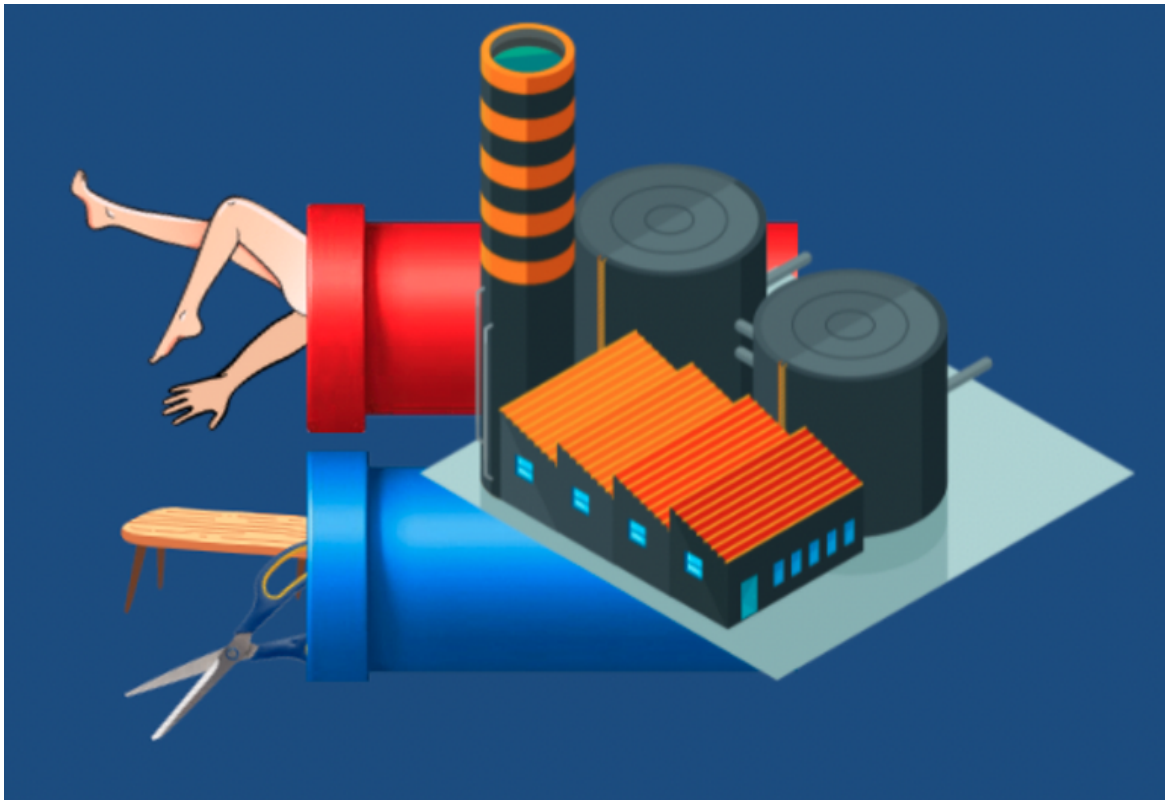
Michael: we've been through all those scenarios. I think we are going to go through them even more in the next slides, but we can maybe start with defining the basics of the models you use for your paper which is capital, labor and output.



Phil: We were talking about GDP before, all those cars and haircuts and so on. Where do they come from?

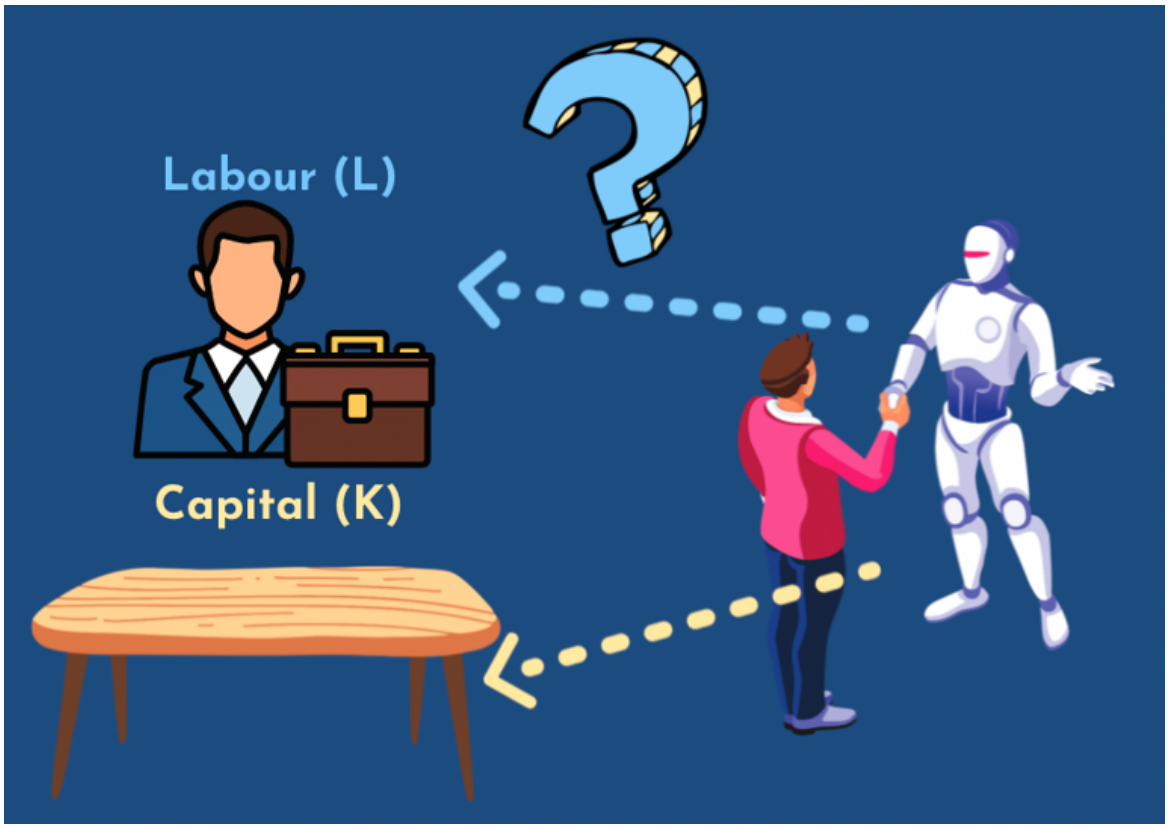
Michael: from desks?

Phil: well, some of them come from a desk. I'm at a desk now. They use a lot of inputs but we can divide them into two categories labor and capital where labor is obviously human input and capital is desks and factories, and the scissors that the barber uses and all of that, and we have a function, F there, which takes in all the labor and all the capital and spits out output, GDP.



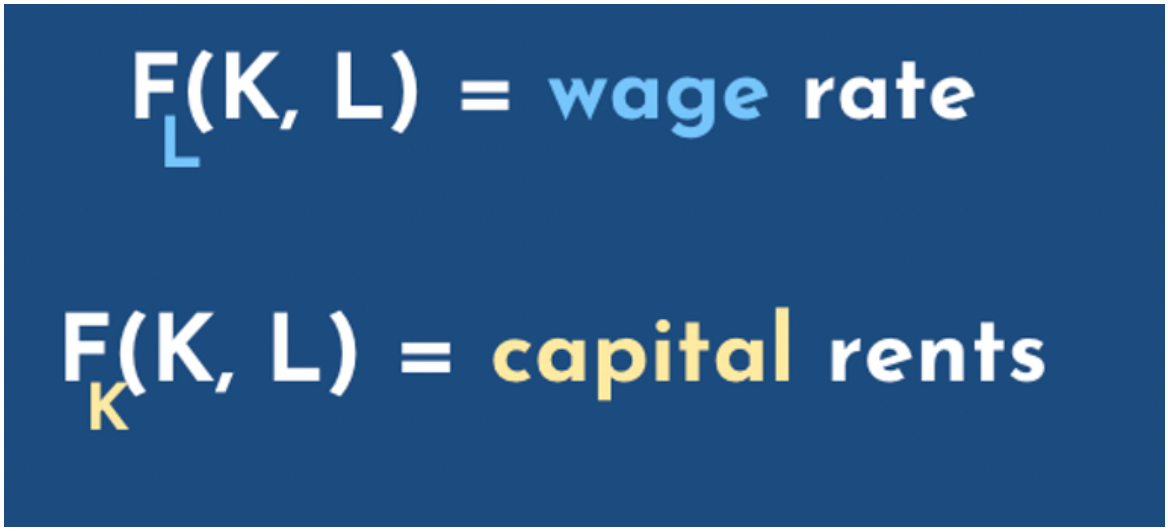
Phil: So you can think of it as like the whole economy is just a big factory with two big tubes going in, the worker limbs and the desks and metal and everything and then the outcome is GDP.

Michael: it's a bit apocalyptic to have humans in tubes... but I get the idea.



Michael: To go back to AI, if AI automates labor and like start to think like humans and does work like humans, we want maybe to consider them labor. And if we're just like building more computers they're more obviously capitals. If they're like servers or so... I guess then the distinction between those two is less thick for AI.

Phil: I agree.



Michael: so let's go further, so for for this I guess I just try to... the little L and K are your notation, which is just a derivative... the partial derivative for the labor and for capital and I guess that's if we're assuming that we're paying people their marginal products. So could you define what's a marginal product?

Phil: hum, sure. The marginal product of a worker is the extent to which output increases holding everything else the same when that worker puts in another hour of work or the amount it falls when they put in an hour less work... not an hour, an infinite decimal amount, but let's say an hour. And in a world with competitive markets, as they're called, the workers hourly wage should be their marginal product per hour in this sense. Why? Well, because if you've got, say, factories around or you have barber shops around, if a given factory or barbershop or whatever isn't paying someone their marginal product then they can just go to a similar one across town and say: "pay me a bit more" and they'll be willing to do so. Things don't always work out that neatly but I think it's a good place to start. So that's the marginal product of labor and why you should expect it to be the wage. By the same token, the interest rate should be the marginal product of capital, so the annual interest rate. Should be, let's say, I have some some equipment that can be used in a factory or has some scissors or something and I lend them... I lend it to the... let's say it's scissors. I lend them to a barber shop. how much...

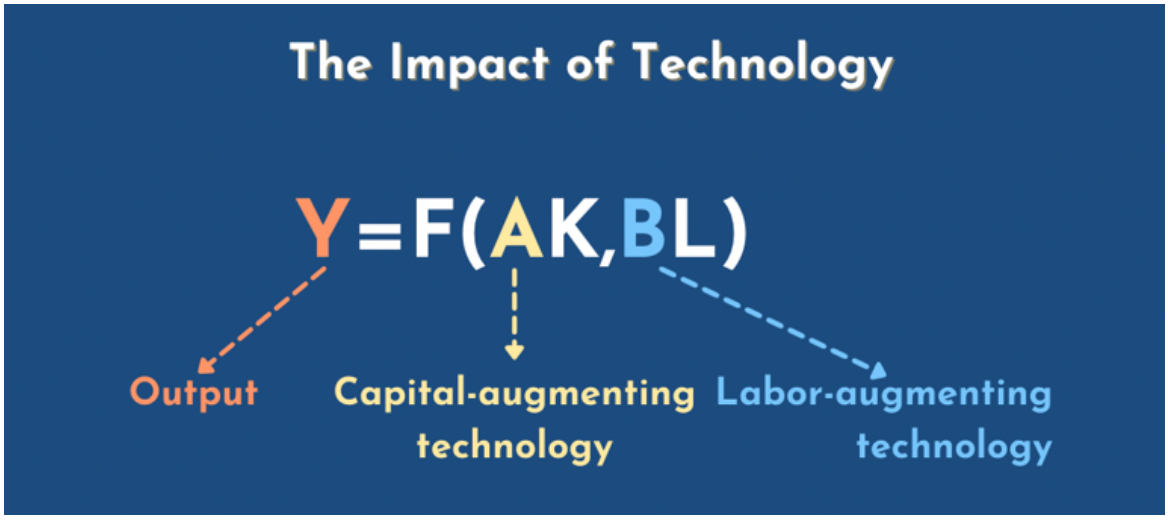
Michael: very nice of you

Phil: well but I'm going to get paid

Michael: oh yeah

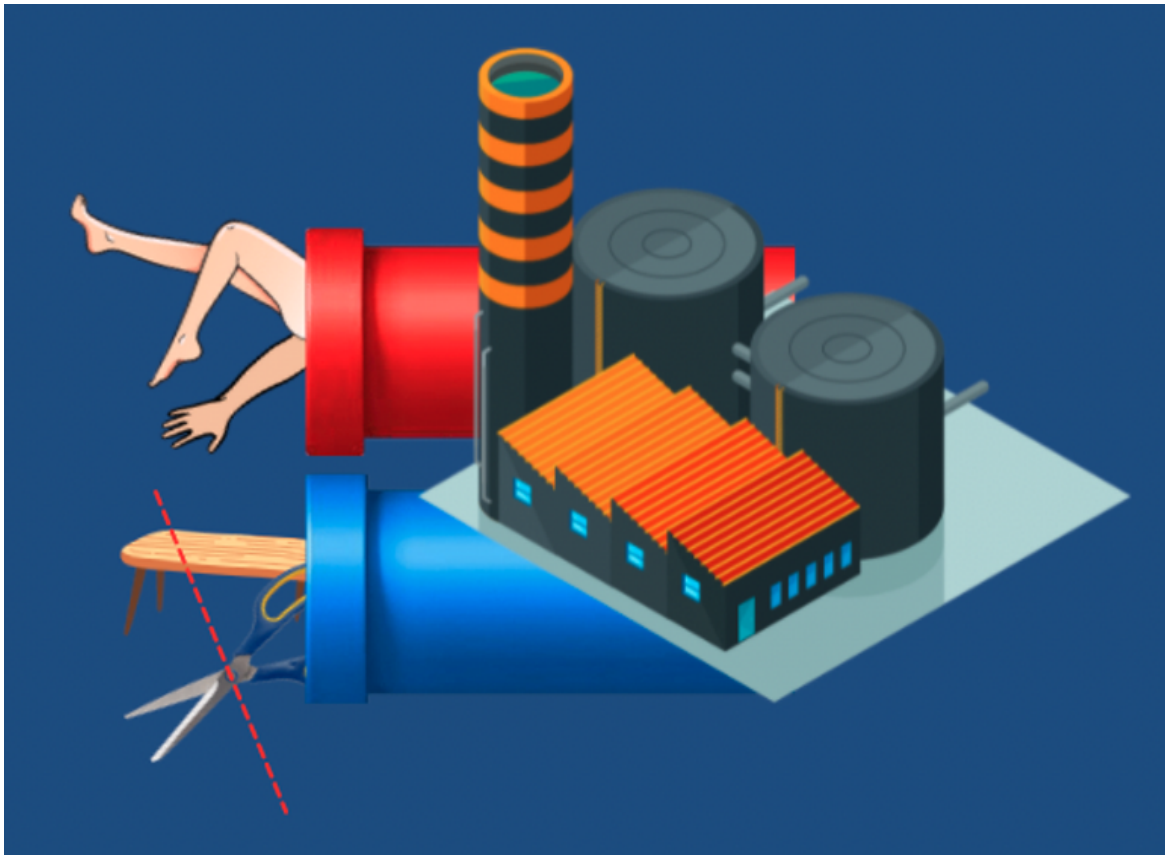
Phil: How much am I going to get paid? The extent to which they they're able to make more money over the next year because they have an extra pair of scissors and so that's the marginal product of capital and that should be the capital rental rate or the interest rate, same thing.

Michael: yeah, so I guess more interestingly is when you start to insert technology into that, and so



Michael: you can either have multipliers for capital, capital augmenting technology or multipliers for labor, labor augmenting technology. How do you distinguish those two individually?

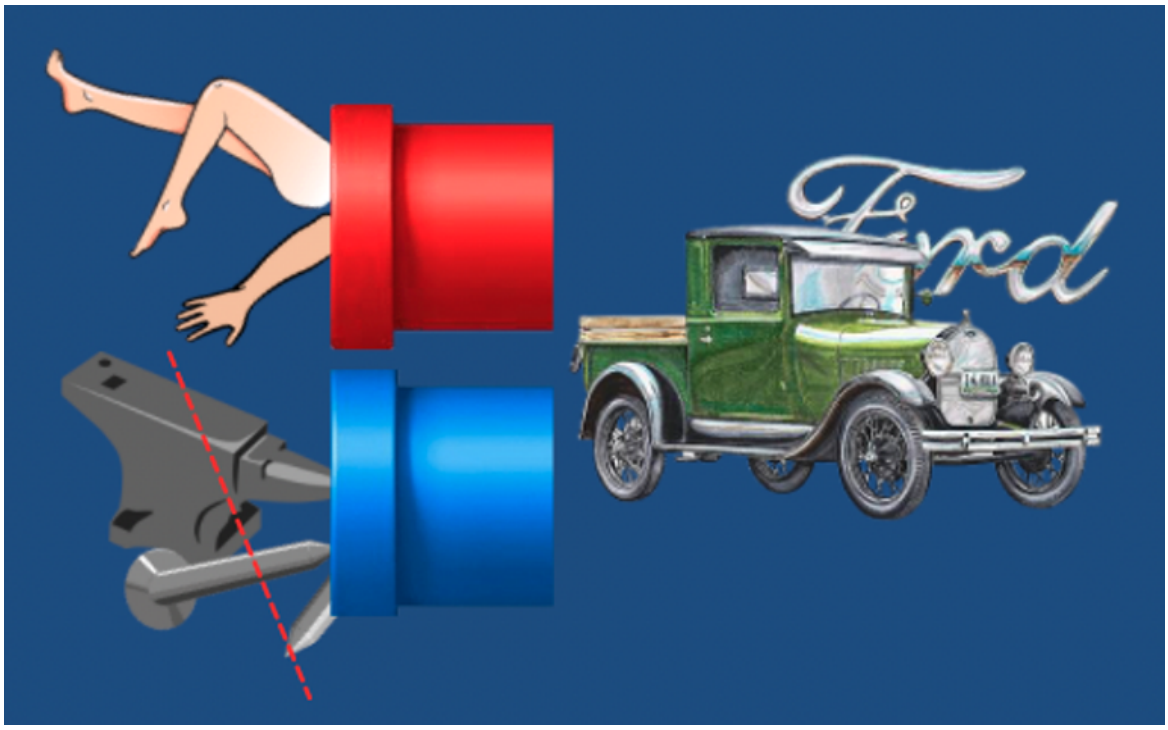
Phi: let's say we come up with a way to re-organize the factory—the metaphorical big factory of the economy which has two tubes going in and it has one tube coming out.



Phil: We reorganize it so that the same amount of output comes out the other end if we use half as much capital going in as before let's say. So you still need to put in the same amount of same amount of bodies but now you can just put in half as many screws and scissors and stuff and the same output comes out as before. Then, that's what we'll call a doubling to A, the capital augmenting technology. A doubling, does that make sense? Because now you only need half as much K to get the same Y, holding the B and L constant.

Michael: But this technology is still a bit surprising: how could we need half less screws?

Phil: oh well this happens all the time it takes a lot less metal now to make a car than it used to back in Ford's day.



Phil: Somehow we've come up with ways to to just shape the structure things... shape the car itself so that the the pieces can all be thinner and lighter and not use as much metal and that's also good for gas efficiency.

Michael: I think for production of materials, or for cars... for products it's quite easy to measure it, I think, in your paper you mentioned that to measure technological development in general, sometimes it's very easy... it tends to be very hard because you can really have like very long-term repercussions. Like I think you said "the importance of the atomic bomb is not well approximated by the cost of the Manhattan project". When you were writing this, what were you thinking about the other consequences or externalities of the atomic bomb?

Phil: naively you might say: "how important (in some economic sense) is a given technological advance?". Well, that's not naive, that's a reasonable question. A naive thing would be to then, say, well, it's going to be reflected in how much people were willing to pay to develop it. So if you've got car companies and they're making their cars and they... you see... that they don't just spend on steel and on workers to put the steel into shape to make a car, but they also hire some engineers to think about how to make the same car with less metal, you might say "well how important is this new design that saved some money when making the car". Well, we can just say, well, how much were they willing to spend on engineers to come up with this?

Phil: That should be something like how much they saved, because if they're saving more than then they spent on the engineers then, yeah, it's sort of like, they have some market power. They could have invested in even more engineers, and they probably would have been saving more money on that margin, so, at first glance, you might think "okay the value of the invention is something at least for those sort of marginal inventions... for things that the company was just indifferent between funding and not funding, those are the ones where the cost is about the same as the extent to which it actually increases output" but for all kinds of technological advances their impacts are not captured by the people funding them.

Phil: They have lots of impacts on the world, and so those impacts are many steps removed from the organization, or individual deciding how much to spend on it. And then they might be totally unforeseeable, and have ramifications to the generations. And it would just be silly

to use this framework very widely, and, as an extreme example in which that sort of framework would break down, you can consider the Manhattan project which introduced atomic weapons to the world. And I don't know how much the US government spent on the Manhattan project, but it's not... you can say that if you're trying to figure out what are the most important ways in which atomic weapons change the world, looking into how much Einstein got paid, it's not going to be... not that much compared to the value of winning a war.

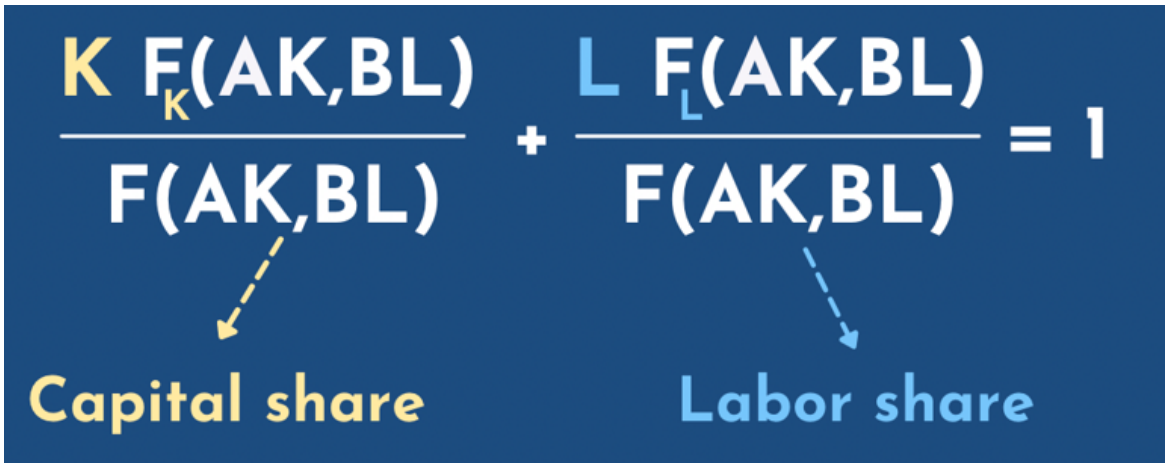
Michael: and, I guess, when I ask you to distinguish the two, it's because when I try to think about what AI could bring, or the internet, or when you people use emails or slack and build new servers, new software, they're kind of things that... they can make computers go faster, and more efficient, and they can also make so... if my computer is more... is faster, the human using the computer will also be more efficient so it will be a labor augmenting technology. In some sense if we have some innovation that enables to make all the computers twice faster, it would both make capital and labor hum, better off, I believe.

Phil: well, there is a difference between making a certain factor, or its owners, better off, and making it more productive. In fact, if this production function F here... if for F, capital and labor are what are called gross complements, meaning that basically you really... if you have a lot of one thing you don't really need any more of that thing you just need the other thing.It's left shoes and right shoes. That's how capital and labor are. Then making a certain factor more productive is bad for the owners of that factor, and making the other factor more productive is good, so if you've got a factory, and you've got machines and people, and you need one person for a machine, and then you suddenly come up with labor augmenting technology, that now means you only need half as many people. You come up with a new training, say for... so the workers can operate twice as many machines at once, then you can just fire half the workers and what happens to the wage? Well it goes way down because no one's really all that necessary you've got this sort...

Michael: you're saying that there's some kind of interaction between those two where if one becomes more productive, or efficient, then you can just remove the other one or remove half of it?

Phil: yeah, so that's in a static setting. Now, when you really think things through, then you have to say well, hold on, if we have all these extra workers lying around, and we can pile up the capital, well at the moment, maybe wages, will go down, but in time we'll end up with twice as many factories as we had before. Because that, well, we'll just accumulate capital and we'll be able to make money. We'll be able to make out... building new factories and hiring all these people that became unemployed, and at the end of that process people's wages should be higher. They should be twice as high in this case because each one of their marginal product will be twice as high, because if they leave now, two machines will be unused, instead of only one, but you just have to be careful between that example you gave about computers going twice as fast. I think if that's all that happens and nothing else changes, but computers get twice as fast, that's a case of capital augmenting technology not labor augmenting technology. But it might often lead to a rearrangement of work processes or production processes that amounts to labor augmenting technology, but in itself it's just capital augmenting technology, even if it makes workers better off because now they can get more done, that's a capital complementing their work.

Michael: it's as having a bigger desk, making people working on the desk better workers... so I think I didn't put all the conditions but, yeah.



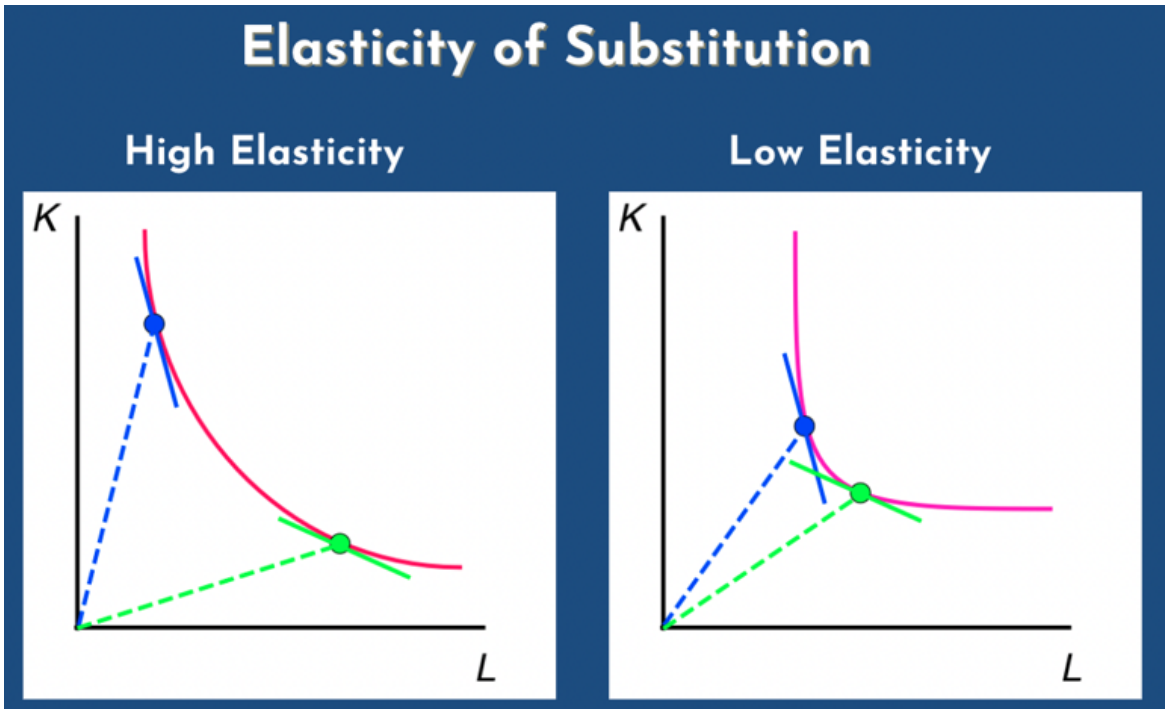
Michael:You already mentioned the wages and capital rents, but another terms are capital share and labor share, which we would also say if they go down or up in our scenarios... and there are certain assumptions that maybe we should not mention because they're mathematical assumptions. We get that those two sum to one. Do you want to explain this formula?

Phil: sure. So we said that each unit of capital got paid by its owner the marginal product of capital each year. Let's say, if we're thinking on a yearly scale, so that's that F sub K thing... it is a marginal product of capital and you multiply that by K the amount of capital and you get that the total amount that all the all the investors took home that year. And then you divide that by F of AK and BL... you divide that by the output, and you get the fraction of total output that got taken home as interest on investments. And then by the same logic, you think about, you know, F sub L: that's the the marginal product of labor, that's the wage, multiply that by L, the amount of labor, that's the total amount that workers take home as wages, divide that by the output and then you get the fraction of output that got taken home as wages, so if nothing's too screwy, those two things should add up to one. 100 percent of all of output got taken home either as capital capital rents or as wages, and in practice I just said the ratio has been about two-thirds.

Michael: and this has been consistent for the last century or something?

Phil: I think more, but the data gets worse the further back you go. But at least for the last century or so. And in recent decades, the labor share has fallen a little bit now. Maybe it's not... instead of being 67 percent now maybe it's 62 percent in the US and UK, and people make a very big deal about this, blame it on Reagan or something, but I think it's... I think the more striking fact is just that it's managed to stay so roughly constant for such a long time, even as the economy has grown a lot and industries have risen and fallen, and yeah...

Michael: and maybe with AI we'll be automating some part of this labor share, and it will shrink a bit, which is in all the scenarios you you will consider further on. We'll see if the labor share stays constant or goes down a lot, or even shrinks to zero.



Michael: Yeah so I think another concept that is central to this literature review is elasticity of substitution, and maybe to understand that... this took me a little some time to understand, so if someone has never... maybe have never heard of elasticity in general for normal products... and maybe you could start with that.

Phil: sure, so the elasticity of one variable with respect to another variable is the percent that the first one rises when the other one rises by an increment. If you're thinking about the price elasticity of demand for some apple, what you're asking is when the price of apples rises by a tiny tiny bit, it rises by a penny ...

Michael: I don't really look at the prices of apples.

Phil: Well but, okay, but you would if they if they got way higher, probably, if it was five times higher and there's someone out there for whom the current price is five times higher than the amount that it used to be in the amount they would be willing to casually buy them at... so there aren't going to be many people who stop buying apples or many apples they refrain from buying when the price goes up by a penny but the question is well let's say, a pound of apples is two dollars, so it goes up by a penny, so that's half a percent, to what extent does the quantity of apple's produced fall, if it falls by half a percent? Then the elasticity is one if it falls by one percent. That allows that the price elasticity of demand for apples is two.. so that's that's the idea and...

Michael: could you give examples of products that are like clearly elastic or where the price elasticity of demand is elastic, and or inelastic, so we have concrete real-world examples?

Phil: Sure, so how about some particular kind of apple? This is always... it's a always a little arbitrary, what you count as a good and what you count as a category of goods, but let's say we're just thinking about pink lady apples. I'd be very surprised if the price elasticity of demand for those were not well above one, because there are very close substitutes, right? On the shelf next to it, so people...

Michael: I guess one intuition for that is when the the price elasticity of demand is high then, if the prices changes people will just go for the substitute and they don't really care about this pink lady apple or something.

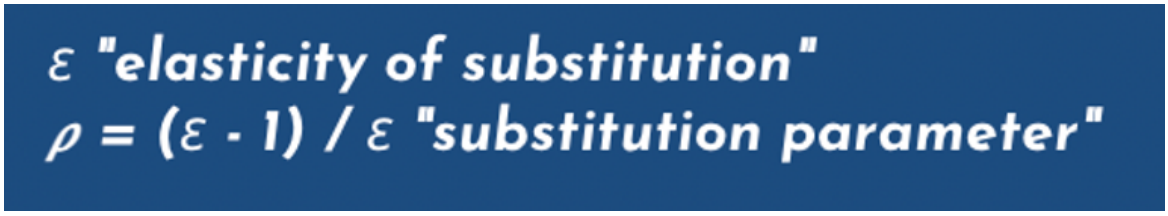
Phil: sure.

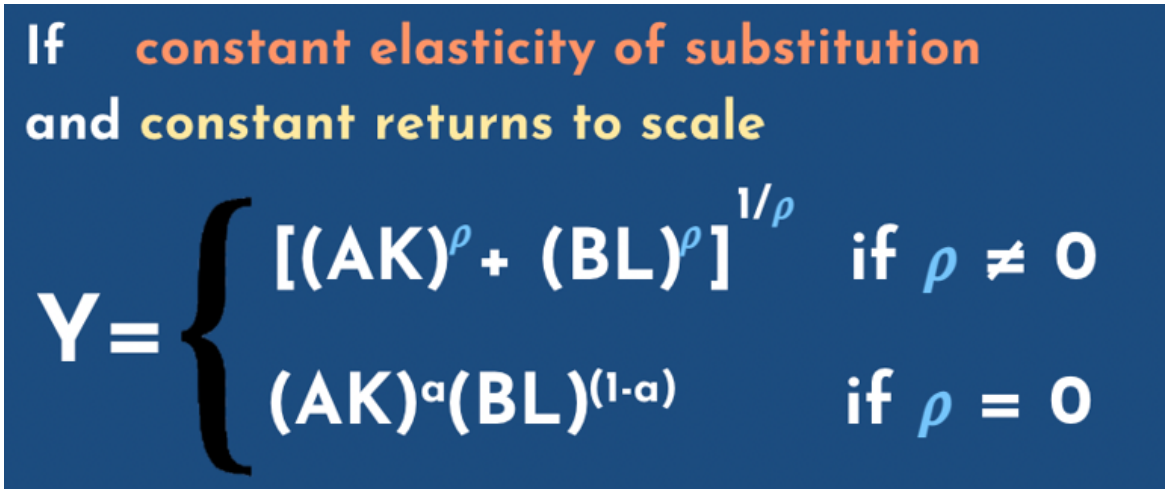
Michael: and how do you translate this concept to, you know, capital and labor, and this elasticity of substitution.

Phil: so a key thing to keep an eye on, is the "elasticity of substitution" between capital and labor in production. And what that is, let's say that the quantity of one of the factors rises by a bit. How much... well, technically it's by how much does the usefulness of a bit more of that thing fall, okay? So, we've got our big factory with two pipes in, one pipe out. Let's say you have one percent more capital going in than you had yesterday but you have the same amount of labor. Let's assume they're complements. They don't have to be gross complements, which is a stronger concept that I mentioned a minute ago. But they're just concepts in the sense that if you have a high capital to labor ratio, a lot of capital per unit of labor, then labor is more useful than if there's a low capital to labor ratio and vice versa.

Michael: yeah, just to to have some clear visual for this. You see in the high elasticity... the upper left point we have high capital and low labor and so if we... this red line or pink lines are isoquant, which is the same amount of output and so here, in this point, if you... as you said... if we increase a little bit of the amount of capital by 0.1, we would increase more, uh sorry, if you oh sorry, if you increase capital then you would decrease a little bit less labor because the derivative is so depending on the derivatives you get how much you'd change need to change the other variables to get the same output.

Phil: that's another way of putting... it's probably a better way of putting it, actually. Well, I don't know. Sure, so if you have one percent, or a tenth of a percent less capital, how much more labor do you need to fill in, and if the two factors are highly elastic then you don't need that much extra labor you can just... you can just sort of fill it in without too much trouble but if there's low elasticity of substitution between them, then if you're running short on capital you need lots and lots and lots of labor to make up for the difference and vice versa, yeah.





Michael: so I think that this will be more clearer in the next slides where we define... so in your paper you have epsilon, which is elasticity of this substitution and... that's going to go from... I believe zero to plus infinity?

Phil: yep

Michael: and then you can divide... you can do epsilon minus one divided by epsilon and this goes up to one?

Phil: yeah

Michael: and I don't know if you consider case over where it's negative and infinite or something it can go down

Phil: yeah

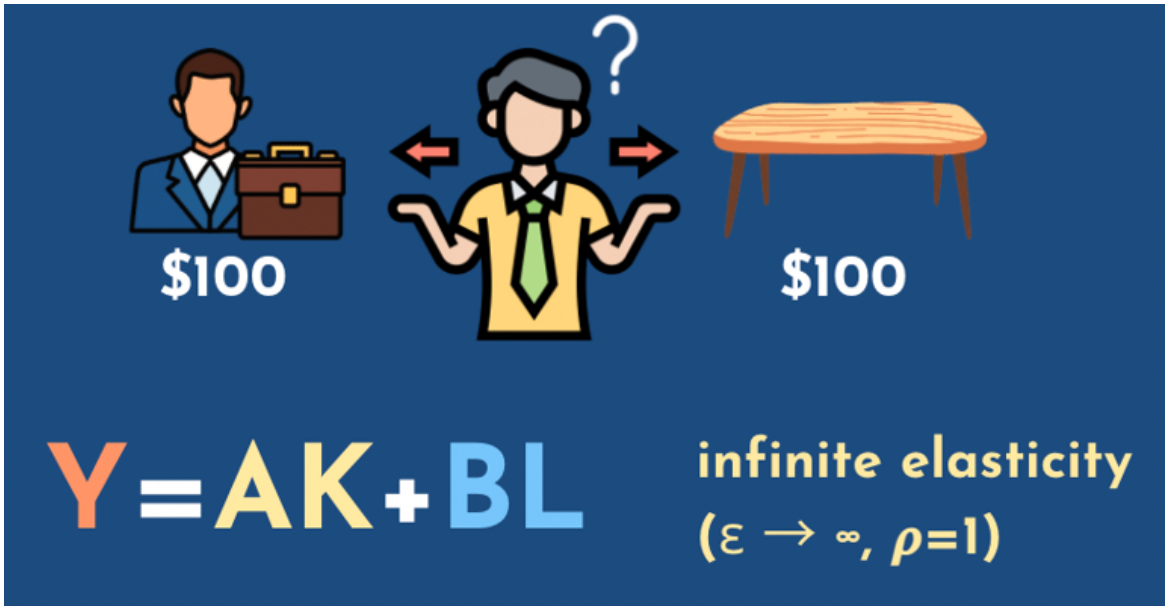
Michael: so yeah do you have intuitions for this parameter rho, or substitutability parameter?

Phil: I think you don't need intuition for all that much. I think all you really need is to remember the following: if rho is one, that's the same as infinite elasticity of substitution, or perfect substitutability, and in that case it doesn't matter which input you have, it's just: you have humans and robots side by side in the factory or whatever, cutting the hair, and you could just swap out one human for one robot and or vice versa. It's perfectly substitutable, it doesn't have to be one for one, there could be some constant ratio maybe. The robot zips around twice as fast, so a robot's as good as two people, but they don't complement each other. It's not that if you have more robots around it's better to have an extra person or vice versa. So that's rho equals one. As rho goes down to negative infinity you get to the case where they are perfect complements, and that's left shoes and right shoes. If you have more of one than the other then you don't need any more of the one that you have in excess, and again it can be at some ratio, so bicycle frames and bicycle wheels, it's not a two to one ratio but it's the same idea yeah.

Michael: the basic intuition with shoes is that you need both, so even if you don't have... even if you have two left shoes and zero red shoes, you would still buy a right one because you want... you need both.

Phil: right and the marginal value of another left shoe to you is zero. It's not just low in that case. It's actually zero. Now an interesting thing happens around rho equals zero, in the middle there. If rho is greater than zero, even if it's not one, so the factors aren't perfect substitutes, they're substitutable enough that they are what's called gross substitutes. They're not gross complements, which is a term I raised before. And what that means is that if you just have a fixed amount of one of the two things, labor, say fixed amount of labor, but you pile up the capital, then output goes to infinity. You don't need labor and, likewise, this labor goes to infinity, but capital stays constant, output can go to infinity. If rho is less than zero, then you need both. And you might think of workers at their desks. You might be a bit more productive as your desk gets bigger. And as your computer screen gets bigger, and all of that, but as that goes to infinity, as the screen gets bigger, the desk gets bigger, your output just goes to an upper bound and it might not be that much higher than the upper bound you have now. At least I find if I had a lot more capital to help me out I don't know what I would even do with it after a while, maybe 10% more productive or something, yeah.

Michael: I think those examples where rho equals one and minus one are very good to keep in mind.



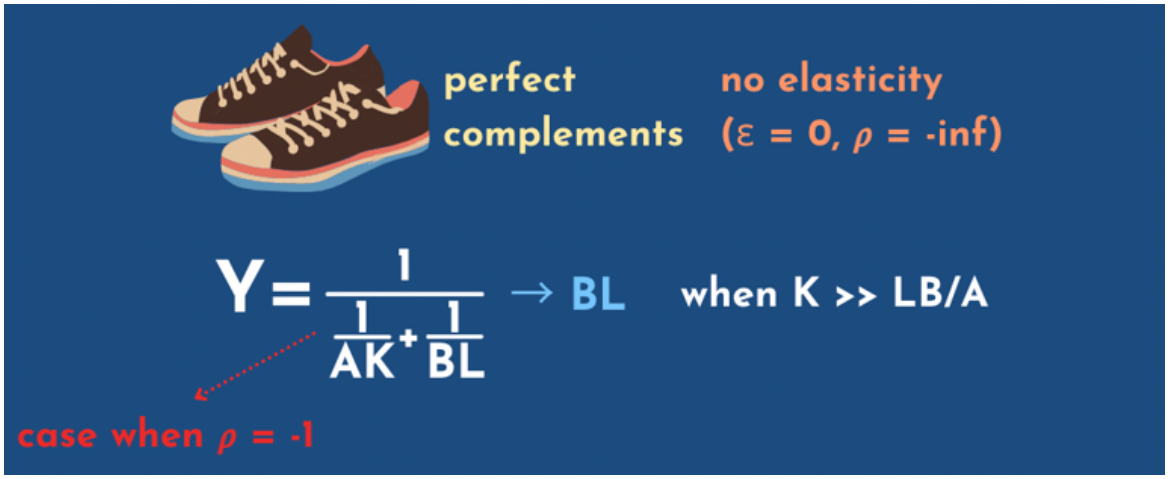
Michael: So with one, as you said, we get a perfect substitute, where you get this formula, pretty simple, where you have the sum of the two, and you could either go for capital or for labor, and as you said, it corresponds to infinite elasticity. So you could just be completely indifferent between the two, capital and humans. I put this diagram with human and desk, because you would only be indifferent between those two if they had exactly the same prices.

Phil: If they were perfect substitutes, yeah.

Michael: yeah, if the price of human labor is too high, you would just go for desk, and the example was minus one... if this example is... choose where you have this fraction.

Phil: no that would be negative infinity

Michael: oh sorry,um, hmm so oh yeah oh that's a mistake because when epsilon tends to zero you go to minus infinity.

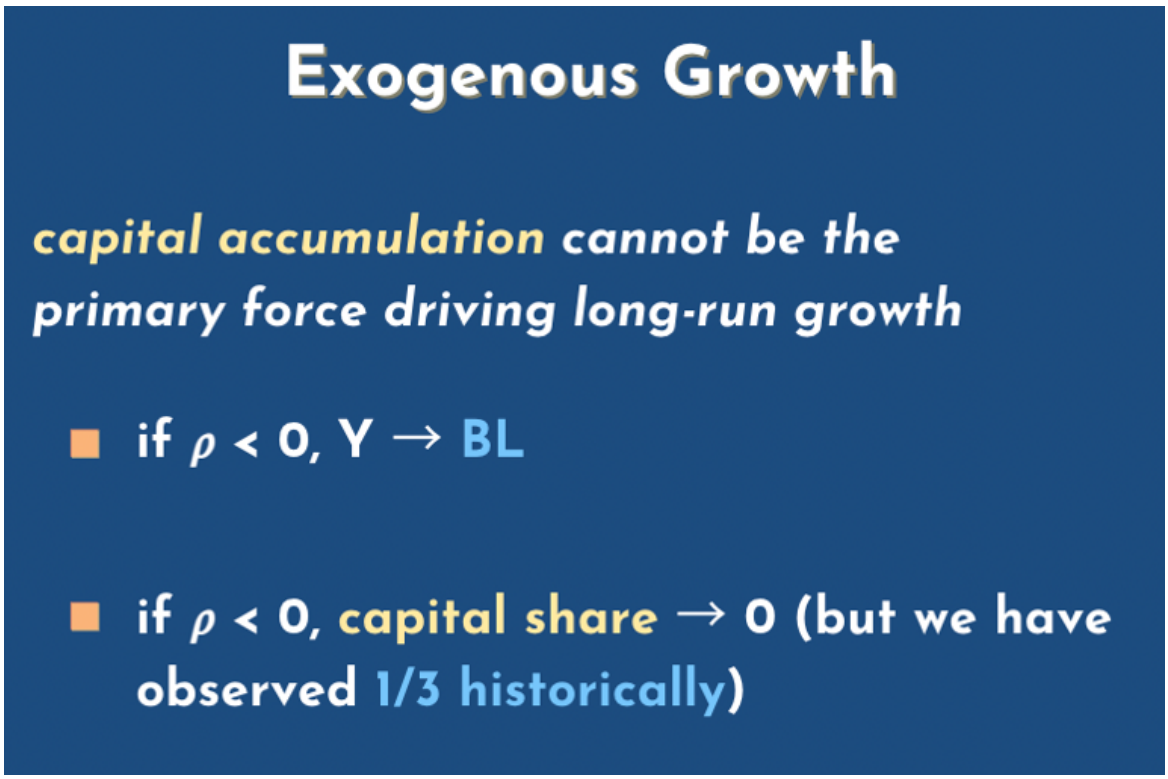


Phil: yeah so there's nothing special about the case of rho equals minus one it seems. There should be, maybe, but I like using it as an illustration of the math because it's kind of easier to wrap your head around than... that funny thing you had a few slides back.

Michael: We have... yeah... so it's not perfect, there's a mistake. This is the case where just rho equals minus one, and you get this simple example where you see when you increase capital too much then you only get it, just the fraction tends to zero and the only thing that counts is labor.

Phil: that's right I think so this is this basic tendency where as one of the factors goes to infinity output just goes to being equal to how much you have of the other factor or linear in how much you have in the other factor. That tendency holds for any value of rho that's negative.

Michael: So I think that's interesting to keep in mind. This example of simple math for when rho is negative... or when rho is always positive. We have this sum and and then



Michael: there's this concept of endogenous versus exogenous growth. I think one line of your paper is "capital accumulation cannot be the primary force driving economic growth", and then you go into different cases for why it's not the case. Can you delve into that?

Phil: Sure, so empirically it seems that, at the moment, before we've come up with... AGI... capital and labor are gross complements. So rho is negative, and what that means is that you couldn't sustain growth just by piling up the capital, because it's like people sitting at a desk and the desk gets ever bigger and you have ever more pens in the drawer and, like, what do you do with them? So as as we piled up the capital we would just get output going to an upper bound and that would be proportional to the number of people. So that's why Y goes to BL, thing you have. And another thing that would happen is that the capital share would fall to zero because it would be like, well, this gets, this is why the elasticity of substitution point is sort of relevant... the thought is: as the number of desks grows, and grows by one percent, then another percent, then another percent, what happens to the marginal product of an inch of desk space? It falls. But how fast does it fall? Well, if rho is less than zero, that is if the elasticity is less than one, the marginal product falls by more than a percent every time. So you have one percent more desks but the rent that they take... that each each desk lender gets, follows by more than a percent.

Michael: okay.

Phil: and so that means that the amount they're all getting collectively is shrinking even as the amount of deaths is rising and they stop investing at that point... or they probably stopped investing but if they didn't, just theoretically, if they for some reason just kept on building desks and letting them out at evergreen, then the capital share would go to zero right, because for every percent extra desks you have, you have less than now, percentage falls by more than a percent okay

Michael: and probably people would also like would you increase of... that they would also have to have more human labor and so this human labor would represent more of this this share of output.

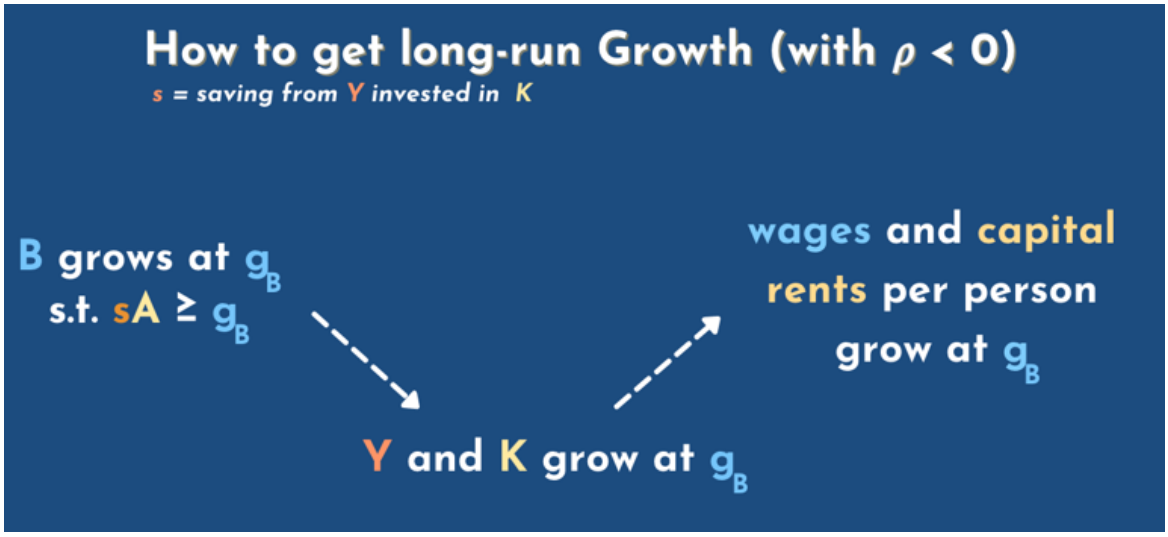
Phil: That's another thing happening. It wouldn't matter after a while, but at least for a while human labor would be getting more productive and that would raise the wages.

Michael: So that's why, like, just investing more capital couldn't sustain growth

Phil: couldn't sustain growth in the long term, and it would lead to a capital share falling to zero but historically we've observed constant capital share and we have seen growth, so capital accumulation can't be what's going on.

Michael: There's some contradiction with empirical facts.

Phil: Not in isolation.



Michael: And then there's yes when we consider negative rho so elasticity less than one this is in your scenario where conditions... can you explain a little bit those scenario, and how to get long run growth?

Phil: sure. The way to get long run growth in output per capita when rho is negative, that is when labor and capital are gross complements, is to have labor augmenting technology. That's the only way to do it. The reason is, that to get growth in output, not per capita but just output period, you need growth in two things. First is either capital or capital augmenting technology. You need more of the second thing. What you need is more BL. You need more effective labor, but so when it comes to growing the stock of effective capital it doesn't matter whether that comes in the form of more actual capital or more capital augmenting technology: either one is fine. But when it comes to growing the pool of effective labor, that can't come from just growing the number of people because then you just have

more mouths to feed. If you want more output per person you need growth in B, that is labor augmenting technology. And so that's the first step: you need be growing at some rate and...

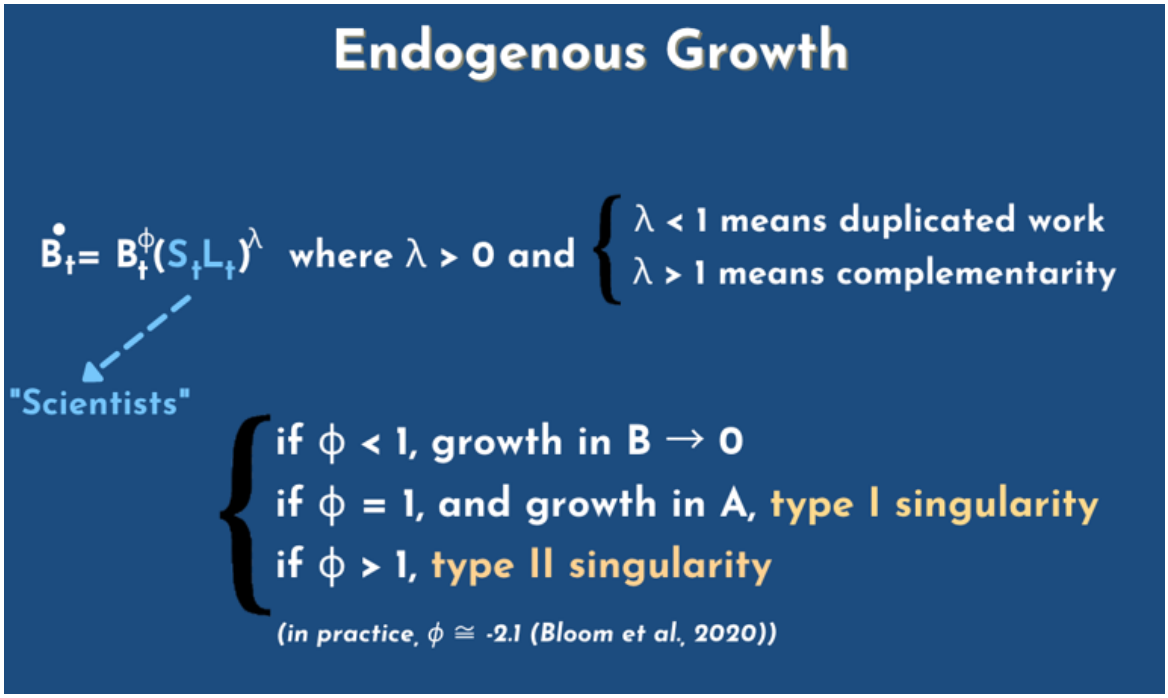
Michael: I thought one thing important... we did talk a lot about product output and GDP, but we never talked about the distinction between people consuming this GDP and or people investing more so maybe you could explain that as well?

Phil: sure, so we talked about all the final goods that came out the pipe at the end of the factory. That's GDP, and I think I might have sloppily said we're just talking about the consumption goods, but I shouldn't have said that, I'm sorry. I should have said "the final goods", and sometimes it's not clear what counts as a final good. It's true, like with the eggs in the cake, right? Is the cake the final good or is the egg the final good that you bring home and use to bake cake. Well in the same way it's a little unclear what's the final good when what you're making is a widget that's then used next year in a factory to make more goods. But anyway let's put that aside and say okay we're just looking at all the final goods produced this year: that's GDP. That includes haircuts and cars but it also includes light bulbs which are not used at home, but are used in a factory to light the place up and let workers work on cars for next year. So that's GDP. Now, remember, we need two things growing to get growth in output per capita. We need growth in B and we need growth in AK. Now let's say A is fixed, so there's no capital augmenting technology growth. How is K going to grow? Well that's what's called saving. That's not consuming all of the output of the factory. Not consuming all of GDP, but using some of it to make more stuff next year, or whatever in the future, so using the light bulbs to light up the factory and not the home...

Michael: or you could just sell sell your stuff? Sell what you produce and then buy more things with the money you make from selling?

Phil: well yeah, but let's not think about... we're thinking about the economy as a whole. If it's the whole world then there's no one to buy or to sell, so... if the savings rate is high enough, then growth in K will be enough to keep up with that growth in B, that g_b, so if B is growing at two percent a year and the savings rate is high enough then you'll have both those inputs to the big factory of the economy, the metaphorical factor growing at two percent a year, and output will also be growing two percent a year. If the savings rate isn't high enough then we'll be constrained by capital eventually and output will be slower.

Michael: right, and at this point you show that output capital growth at the same rate as the labor, superior to g_b, and you get as well wages and capital rents per person growing, so that would sustain long-term long-run growth.



Michael: You made the distinction between endogenous growth and exogenous growth... so could you maybe explain the distinction?

Phil: sure, so as noted we need growth in B, in labor augmenting technology, to get growth and output per capita. We've observed and hope to keep getting... does that growth come from, well, if you're not really going to model it, but you're just going to say, well, somehow B grows at two percent a year and we'll think about the rest of the economy, assuming that's what happens: that's called exogenous growth. But if we're going to actually model where that growth in B comes from we call it endogenous growth. And there's a certain class of models that I think is most plausible, sometimes called semi-endogenous growth models, where growth comes not only from scientific research and applied R&D, but more kind of further down the pipeline, but it comes from people working. But you need growth in the number of scientists to maintain growth and output. You can't just have a constant number of scientists leading to exponential growth and output: you need growth from the scientists and where do growth and sciences come from? Population growth and growth in education—we're not going to model that, we're just going to say that comes out of the sky. So that's why it's called semi-endogenous, but it's it's more endogenous than exogenous, you're at least kind of explaining...

Michael: so endogenous is "comes from inside" whereas exogenous is "from outside"?

Phil: basically, yeah

Michael: and without going too much into the math, but if you have this growth in B that comes from scientists... scientists are a fraction of the labor force that work on, maybe more labor augmenting technology, and from this paper we mentioned, "Are ideas getting harder to find?" we can see that sometime scientists can collaborate and do great stuff and be complementary, or they can step on toes and you need to invest more in research to get more done. So is that the intuition between the parameter lambda, or?

Phil: in this model, lambda ends up not mattering much for the model but, yeah, lambda equals to one would mean that if you double the number of scientists at a given moment, you double the rate at which they come up with increases to B. Increases in labor augmented technology, and if lambda is greater than one then the relationship between

number of scientists at a given time and advances in B is more than likely doubling the number of scientists. It leads to more than a doubling in the speed at which they crank out the increases to labor augmented technology, and if land is less than one it's the other way doubling the number of scientists... less than doubles the contributions in... intuitively we would think that lambda is inferior to one. We have diminishing returns in having more scientists or having more labor in science. I think that is what we tend to observe though. I don't know if that's because of duplicated work or just because you're going to choose the best if you're only going to have a few scientists it'll be the if there's only a little bit of research funding it'll be the best ones and then as you dip ever deeper into the dregs you might double the number of scientists but you won't double the number of IQ points.

Michael: and I guess the other kind of parameter for this model is kind of this power to this labor augmenting technology, this parameter phi... where depending on where it is you can either have something that stops the growth in labor augmenting technology, or leads to type I or II singularities. I have a hard time kind of understanding this parameter. What's the intuition behind it? And you have mentioned that it is empirically at minus 2.1, or at least that's one estimate. Do you agree with that estimate? If you remember... if you remember it? Or can you explain what's the number?

Phil: I'll explain what it signifies, and then comment on it. I call it the research feedback parameter. If it's equal to one then we have purely endogenous growth in the sense that we don't need any increase to the number of, let's call them scientists, to maintain constant exponential growth in B, labor augmenting technology. So that's positive research feedback. What's going on is, sorry I should have explained this first: if phi is positive we have positive research feedback meaning that the inventions we've already come up with, will help us in developing further advances in labor augmenting technology. So we have computers, and those help us in our projects to develop the next big thing after computing. If phi is negative then that might still be true. Inventions in the past help us with further technological advancement but there's this other factor which outweighs it, and it's the fishing out effect: it's that given that when we've already invented something as great as a computer it's harder to invent the next thing, because the next thing is even more complicated and that wins out even though we have computers. We move more slowly with the same number of scientists because things have gotten so much harder and so... if phi is negative with a constant number of scientists, things just get harder and harder and, the level of technology rises to an upper bound then it gets stuck. What you need to maintain growth with phi being negative, is you need to sustain growth in the number of scientists. And, as you mentioned, there's a paper from... well it's a few years in the making but just published last year, "Are ideas getting harder to find" which tries to estimate phi and finds that phi is quite negative.

Michael: You need to invest much more money or people, more labor, to get these new advances in Moore's law or other scientific adventures.

Phil: yeah, the key issue is that you need more people, if you just needed more capital then more money just bought capital equipment then actually you could sustain growth and output per person, even without the scientists multiplying because you could just pile up the factories to do the science and the real issue is that you need growth, and growth in labor input, but we're trying to get growth and output per capita.

Michael: so I think now at this point, we kind of have an idea of this parameter phi, this research feedback. Maybe we can go over... more generally the different kind of scenarios you draw down the line.

Phil: you asked me what I thought of the estimate for my argument

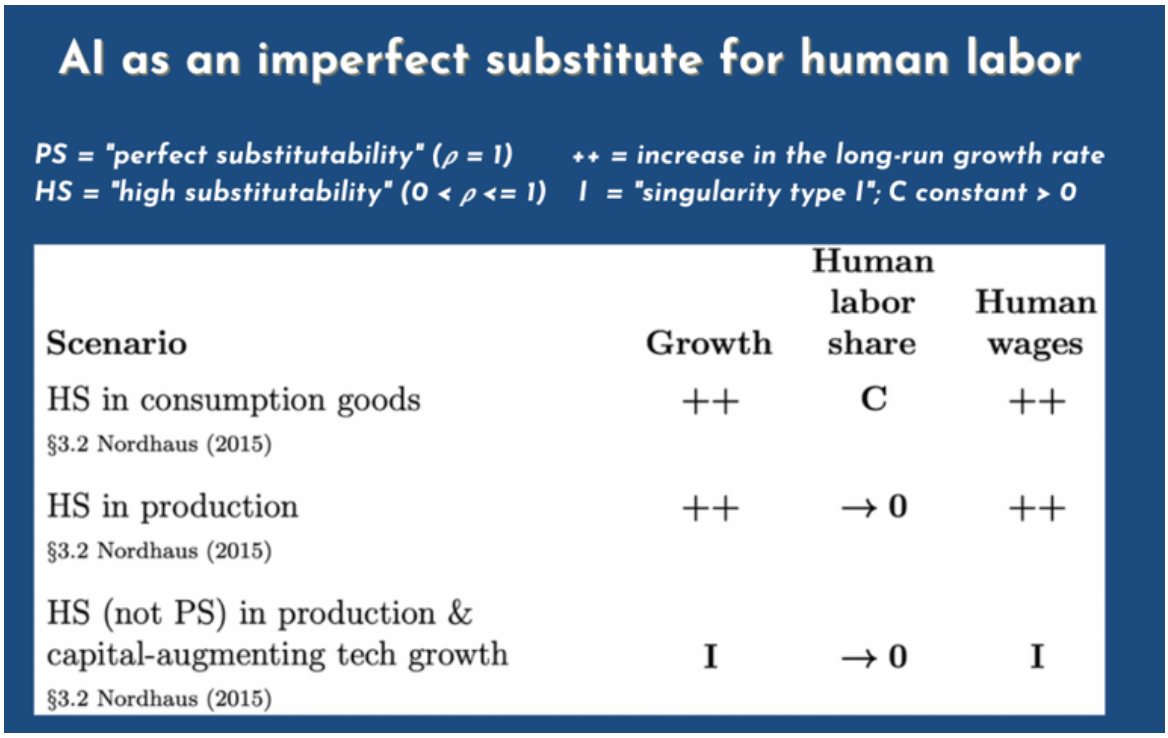
Michael: oh sorry

Phil: what they do is they sort of look at a bunch of different industries and estimate phi in those industries, but actually it varies a lot across industries, and it's just for the economy as a whole that it seems it's negative. If I remember right, the only industry in which it's

positive is actually in Moore's law. I've forgotten about this when... oh no, maybe there's two, but anyway. One of them is in Moore's law, so it seems that the better computers we have aht the faster chips help us make faster chips, buy more, then it gets harder to make faster chips because we would pick the low hanging fruit and that seems very significant especially when we're trying to make projections because if a kind of growing fraction of the economy is being done by computers, or whatever, then maybe the positive feedback loop part of it will end up taking over, and so it's sort of a mistake to think that the composition will be the same and so you have this average negative two. That's constant in the future.

Michael: so if for AI and Moore's law, it's positive, and for other scientific endeavors it's negative, and in average negative, it doesn't mean that we couldn't get labor, and we need to have it as a bigger term because most of it would come from AI or something.

Phil: exactly, and this is something I haven't really paid attention to when I talked about this on "here this idea", or I haven't really thought about too much until recently, but it's only slightly positive and it's zero... it is very much in the error bars. So I think more research on actually confirming that there is this area of positive research feedback would be valuable and picking it apart right, because it might be that it's negative in the domain of AI even though it's positive in the domain of just fitting more transistors on a chip and so that's all I have to say about that.



Michael: Cool. I think we can just skip the the next one and just go on the different scenarios. So the HS means high substitutability, PS perfect one, so these are the different values for rho, our substitutability parameter, and then I think... plus is when it increases in a significant way, so the type I, not a significant increase but... the labor share could go to a constant C, or to zero. And I is a singularity of type I. So... what's the intuition between those kind of three scenarios?

Phil: okay well we skipped over the first scenario where we talked about different kinds of consumption goods. That was on the previous slide, which we skipped, and I think it is maybe a bit of a digression so I won't say anything about the first line. In the second line we're dealing with a model in which you have that factory of the economy with the two pipes in and the one pipe out, but now, instead of rho being negative as it is now, and has been

positive, it could be between zero and one, or it could be one, but either way here's what's going to happen: the growth rate is going to accelerate. It's going to rise because now growth isn't bottlenecked by growth in B anymore. Once rho is positive, you don't need growth in AK and growth in BL, you just need growth in one of the two. It could be either one. And, in particular, it could just be growth in K, so if you just have a high enough savings rate then it'll... then you could have a growth rate just, you know, if A equals one, which actually seems about reasonable right now, then let's say we save 12% of our output every year then what's the growth rate going to be? It's just going to be 12% in the long run.

Michael: because we're just saving all that and invest that in capital.

Phil: yep, so this is 12% more every year... what happens to the human labor share, or the labor share: it goes to zero because the number of effective humans is staying the same but you have this capital that can

Michael: massive pile of capital

Phil: massive pile of capital that is substitutable for the labor, so it's just more and more robots, same number of people. So the fraction of payment going to the people is just going to zero. What happens to wages? Well in this case actually it depends on whether labor and capital are perfectly substitutable, or just highly substitutable. If they're perfectly substitutable, then wages should just stay the same. So wage growth should just stop if there's no labor augmenting technology, and if there is labor augmenting technology, well the wages keep on rising at the rate of labor augmented technology growth because you have the robots and workers side by side and they're very substitutable... or if they're perfectly substitutable. And so the explosion in the number of robots doesn't affect the productivity of the people beside them so their productivity stays the same. Wage stays the same, unless they get more productive, in which case the wage rises if they are... if a labor in capital... only become highly substitutable though... then because you have this explosion in the growth rate because saving is enough for growth... so the growth rate rises to 12 percent or 20 or whatever, you also have an explosion in the number of robots or the amount of capital that's substituting for labor. However it's doing it, and with rho less than one, even if it's still positive, because we're insisting on its high substitutability but with rho less than one having more of that capital around raises the marginal productivity of labor, does that make sense? So you have you have wages rising and the wage growth rate rises

Michael: so it's a distinction between perfect substitutionability and still suitability but not perfect, and depending on that parameter you would have either the human wages so... we're talking about the second line right?

Phil: yeah

Michael: yeah but if it's not perfect then you would you would have the wages growing.

Phil: yeah

Michael: tthe most interesting kind of scenario is the third line where you have both wages and growth on a type I singularity. So for me it's kind of hard to understand why... how a human would behave with an exponential growth... an exponential wage, what would they buy? What would they do in this scenario?

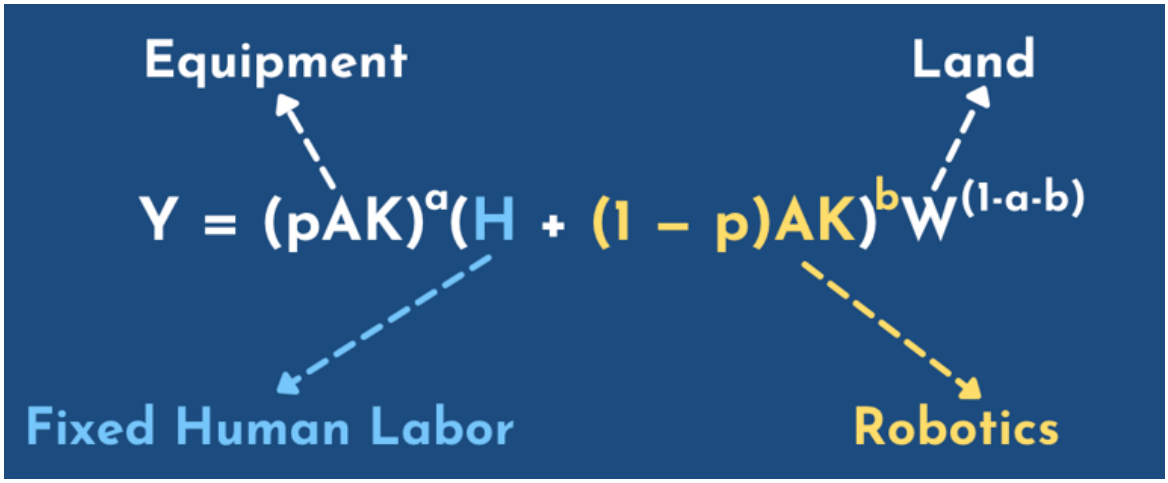
Phil: it would have to be super exponential—we already have exponentially growing wages.

Michael: right, it's super exponential, so that if the growth rate grows exponentially

Phil: yeah, it is hard to imagine though I think it would have been hard to imagine in the past. Like someone in medieval times might have said "what would we do with with a million dollars to spend every year or something on ourselves. But we can have a big castle, but then what a bigger castle?". As time goes on, as people get richer, new products are

developed that means it gets back to something we were talking about before: new products get developed which are, you know, people enjoy buying or spending their money on. I don't know what the products of the future will be but I wouldn't push past people to come up with ways to waste a lot of money and spend a lot of money.

Michael: I think it's interesting to just skip. I think we got the intuition for kind of this imperfect substitute scenario. I think it's interesting to go over...



Michael: if you talk about your medieval guy, at some point the parameter of land was very important so I think in this model... I think it's adapted from Hanson, a paper from Robin Hanson, and you have this parameter of land, and you also have robotics and a fixed amount of humans... so I find this very interesting. Could you maybe explain a bit?

Phil: sure. So in the third line from the previous slide we got a type I singularity because we had capital augmenting technology growing exponentially, and at the same time capital was growing exponentially, so you have this exponential thing growing exponentially, and capital... enough capital, alone is enough for growth. Here, we're looking at a different model in which human labor and capital are perfectly substitutable... in which you can have capital augmenting technology growing, and you can have capital growing, but you don't have a singularity. You just have a boost to the growth rate. And the thing that's different about this model is that now labor and capital aren't the only two factors of production: you have a third thing which you sort of need as well, and that's land. The amount of land is fixed and...

Michael: is the fixed amount of land an important consideration? Do you think it is a reasonable assumption for like, we're doomed to stay on this planet and we cannot build gigantic buildings, so we have this fixed land? Do you think it can be relaxed? And you can think about making buildings with more land in the building? Or colonizing new planets or something?

Phil: I don't know whether we'll ever ever be able to colonize space. It seems very hard. But at the same time we just went over the course of a century from horses and buggies to putting people on the moon. But to maintain exponential, sorry to maintain the singularity you would need exponential growth... in... actually I don't know if that's true. What do you need exponential growth in? Land maybe?

Michael: Hanson maybe thought about this?

Phil: Well, no, this isn't Hanson's model actually. I sort of adapted it from Hanson's paper but Hanson just sort of got rid of the singularity by making output less then... decreasing returns to scale, in labor and capital, whereas, before we didn't talk about this but the other models assumed constant returns to scale and labor and capital... so the idea there is if you just double the number of factories and the number of workers working at them, what happens

to output? Well, it just doubles, right? That kind of makes sense. But no! Because we forgot about the land. If we just had twice as many factories and twice as many people, but the same amount of land, then you have to start putting the factories on land. That wasn't so suitable for a factory, and output wouldn't double. It would rise by a bit less, and that becomes more and more of a problem as the ratio between the amount of land and the amount of other stuff, capital and labor, as that falls. But anyway, so Hanson just made output decreasing returns in labor and capital. He didn't specify land. So he just had pAK to the a, times that other thing to the b, where a plus b is less than one, and I just spelled that out a bit more by introducing land and having them add up to one again, having the exponents add up to one which is more conventional.

Michael: I think it makes sense for me.

Phil: But I was going to say, if we colonized space, then even in the best case scenario land would only grow cubically, because it'd be a sphere of the earth going outward

Michael: and we need exponentially

Phil: so I don't think that would be enough for a a singularity in practice but, anyway, again, that would be kind of probably pushing the model beyond.

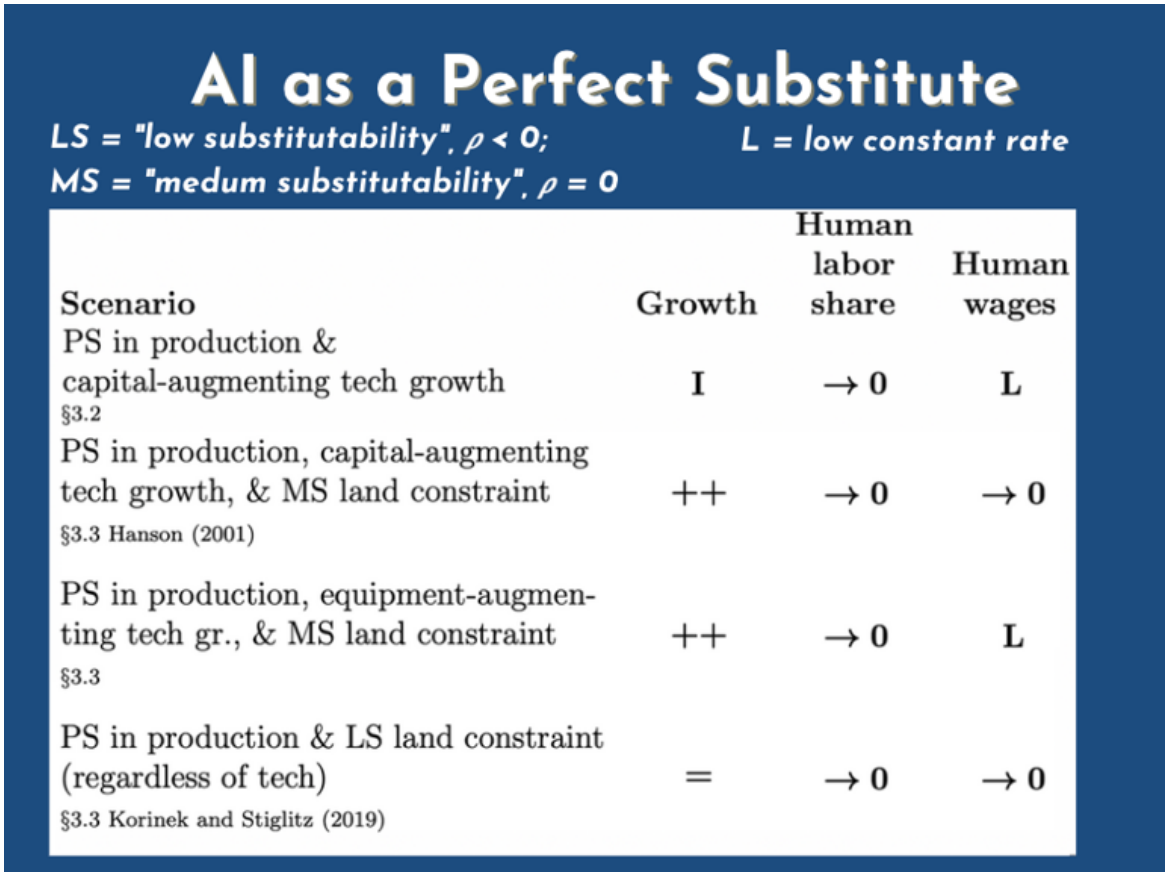
Michael: I guess if you if you really want to push it, you could say that... if you want to test it, what you need is not land... it is kind of resources and energy at the bottleneck of most physical processes, and to get energy you could do a Dyson sphere or something on the sun, and maybe the next centuries or thousands of years you would have even subtler ways of using energy and improving the rate of creating energy.

Phil: so there is that limit of just perfect efficiency.

Michael: so here I think what I understand of this model is... when instead of considering labor augmented technology we have this fixed human labor, and then we can just spend our capital to buy more labor directly, so that's the fraction one minus p that we spend on capital, that can be used into labor and so that that's a perfect substitution. That what we see is an addition between the two and so we can just spend a lot of capital to get more growth and then we just invest more in capital, is that the basic intuition?

Phil: yeah that's right so again in this model labor and capital together are not enough for growth. there are enough... okay so it's sort of right on the edge. I mentioned what happens if rho is positive. What happens if rho is negative in this case? Rho is zero exactly, and what that ends up meaning is that whereas you would otherwise have a type I singularity, with capital augmenting technology growth, and perfect substitutability between labor and capital, here you just get a growth rate increase. Iit's sort of a funny little edge case but when rho equals zero then it sort of bumps down a singularity to a growth rate increase.

Michael: Okay, still interesting, I think it's an interesting equation to think about. So yeah I think the one we get after is the different scenarios. I'm not sure if it's from this equation. I guess, we get those different scenarios. I'm not sure how much we should delve into all of them or if there's one that is especially interesting to you.



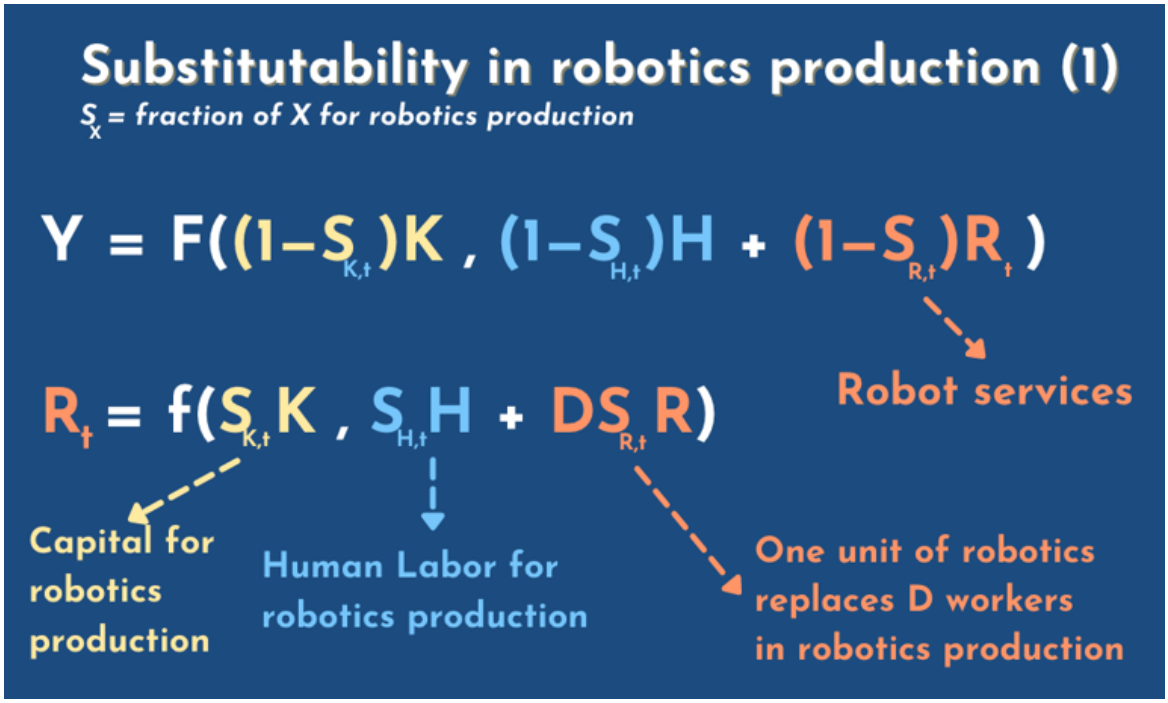
Phil: I don't know. We were just talking about the one on the second line. I guess earlier we were effectively talking about the previous line. We talked about the perfect substitutability in production. There's no land constraint, and you have capital augmenting tech growth, but you get human wages stagnate, and you get that type I singularity.

Michael: If we have no land, then we have type 1 singularity growth, and wages are infinite.

Phil: If we have no land constraints. Quickly the difference here is that instead of having capital augmenting technology growth on the whole we only have technology growth in the the capital that complements labor so we only have ways to use metal more efficiently in the machines that the robots use in the factory... the robots or the humans, but when it comes to the robots we can pile up the robots but we don't get more efficient at making them and they don't get more productive. In that scenario human wages don't fall to zero, because you have kind of two things balancing each other out. One is that we can make more and more robots, which are competing with workers, but on the other hand we have more and more equipment for these robots, and workers to use, which raises workers' wages. But the robots lower the workers wages, but the equipment that gets more efficient raises them, and those cancel out, and the wages just stagnate.

Michael: So I think that's an important distinction. How do you distinguish equipment from robotics? Robotics can help for human labor whereas equipment is just to help humans?

Phil: Robots do labor, and equipment complements labor.



Michael: Interesting. You're talking about what would happen if robots could do the work to get more robust. So you're pointing at somehow the self-improving robots in this model.

Phil: They're not self-improving in this case, they're just self-replicating. In the model you have on the slide and I think the main contribution of this model... this is the Mookherjee and Ray model, is to point out that all the other models we've talked about so far let robots build everything. So they're substituting for labor, whatever we're trying to make, but the important thing is sort of not whether they can do everything but whether they can make other robots. Because, no, the way I should put it is: even if they can do everything else, as long as humans remain necessary in making robots, then humans will still maintain a positive labor share. And so we can we can think about the implications of substitutability in robotics production in isolation instead of in the economy as a whole. I don't think this adds all that much to be honest but it's worth...

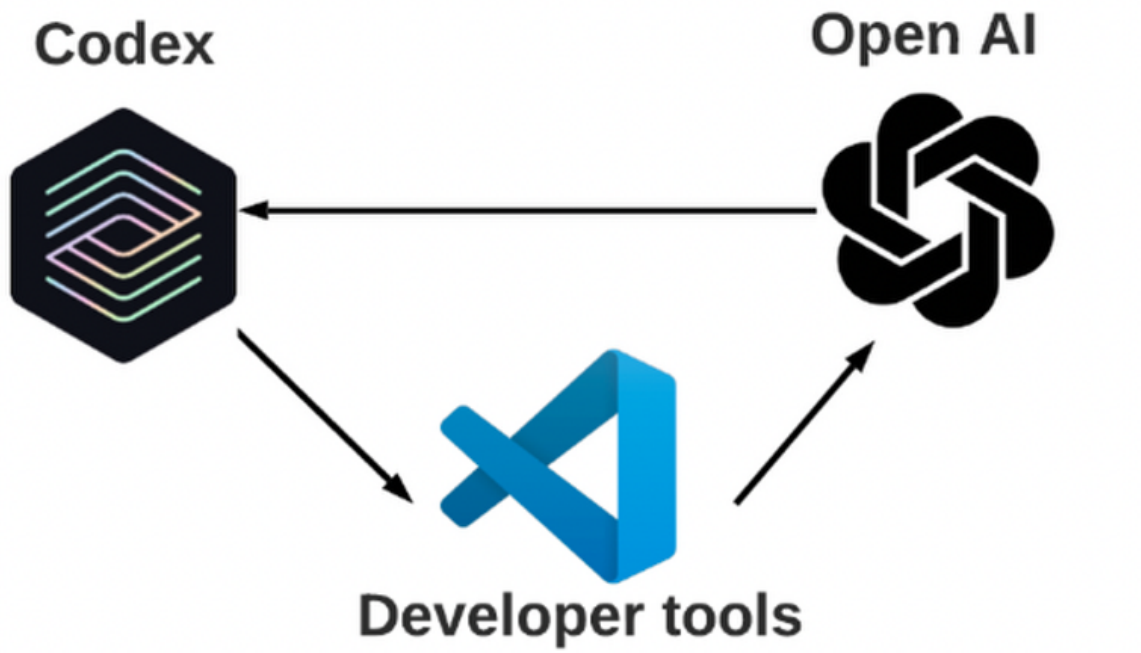
Michael: what doesn't add all that much in these equations?

Phil: I guess it just... the question is "how can people maintain a positive labor share?", if that's your question? "How can people remain necessary for production?" Well, what matters is people being necessary for the production of something that is necessary, or something that people want, and won't just use robots to do instead. And that can be through... a lot of people have noticed that there are a lot of jobs that we prefer done by humans, so some artisanal bread or artwork, or care home work. There might just be some kinds of things we want humans to do, and if we wanted enough we might use our newfound wealth with all the growth that AI brings to just hire lots and lots of people to give us massages and serve us their homes and make us artisanal bread and everyone might end up doing that and you have a positive labor share or growing wages, or whatever it is that you're keeping an eye on. Whatever variable... and one of the ways that people could remain necessary is by making robots... if people are necessary for the creation of robots, but I don't think that's all that interesting. It's just a specific case of a more general observation that we already knew about.

Michael: so you're saying, essentially, that if you want to keep humans, having good wages and having a substantial fraction of the labor share, we should have humans doing something useful and then they could be paid by this amount of growth we have, we could

give it back to the humans doing something useful and humans doing robotics is a special case of this.

Phil: If humans are necessary for doing robotics, and they're at least as useful as robots or complementary I think.



Michael: so I guess if this model is not very interesting, for me, one thing that's has happened in the past in... I think this year, was this company OpenAI releasing something that enables humans to generate code with their model. So you can just start coding and it will auto complete your code with their model directly, and I've used it a lot, and every time I use it, it generates useful code that I can read and reuse myself, and if you think about the system "developers using this, even OpenAI developers, employees plus the model", when we write more code using their model, we can then use this code to improve this robot, or this AI software, and so we're getting in a sort of closed loop where humans are kind of useful but they're not the only part... useful for making code, now AI is also accelerating this.

Phil: yeah but that was already true before, right? We had capital helping out, with the process of improving AI. What will really be the qualitative shift, will be when you don't need the human in the loop. When it can just work on itself without any human

Michael: and if you make the... if you make the human faster. If you make the human writing software, or AI code faster by helping him write more code with AI, would it affect? Do you think, intuitively, it will like lead to faster growth and a type I singularity or, and your view doesn't really change that much in your models?

Phil: again you keep calling them mine: they're not mine.

Michael: oh sorry, in the literature reviews

Phil: I'm not claiming any credit. Okay, so that is an interesting thought, that would be different from what I said. So what I said was that, would it just be capital, that's we already have capital complementing people doing stuff, but this would be a case in which capital is not substituting for labor but making labor augmenting technology grow. So the humans are getting more productive. One programmer can do what it used to take two programmers to do, and I think the model.

Michael: I think this could go into labor augmenting technology. If you're coding faster you effectively maybe...

Michael: I think the the next ones are are essentially different scenarios. I don't know how much we want to go into these ten different scenarios or something, what do you think?

Substitutability in robotics production (2)

Scenario	Growth	Human labor share	Human wages
HS in final good production, HS in robotics production §3.4 Mookherjee and Ray (2017), Korinek and Stiglitz (2019)	++	$\to 0$	L
HS in final good production, LS in robotics production §3.4 Mookherjee and Ray (2017), Korinek (2018)	+	C	C

Phil: I don't know. So these these two lines right here are about the thing we just talked about right now, about substitutability in robotics production. In particular... I don't have much more to say about them than I already said, but for later slides, I don't know.

Growth impacts via impacts on savings

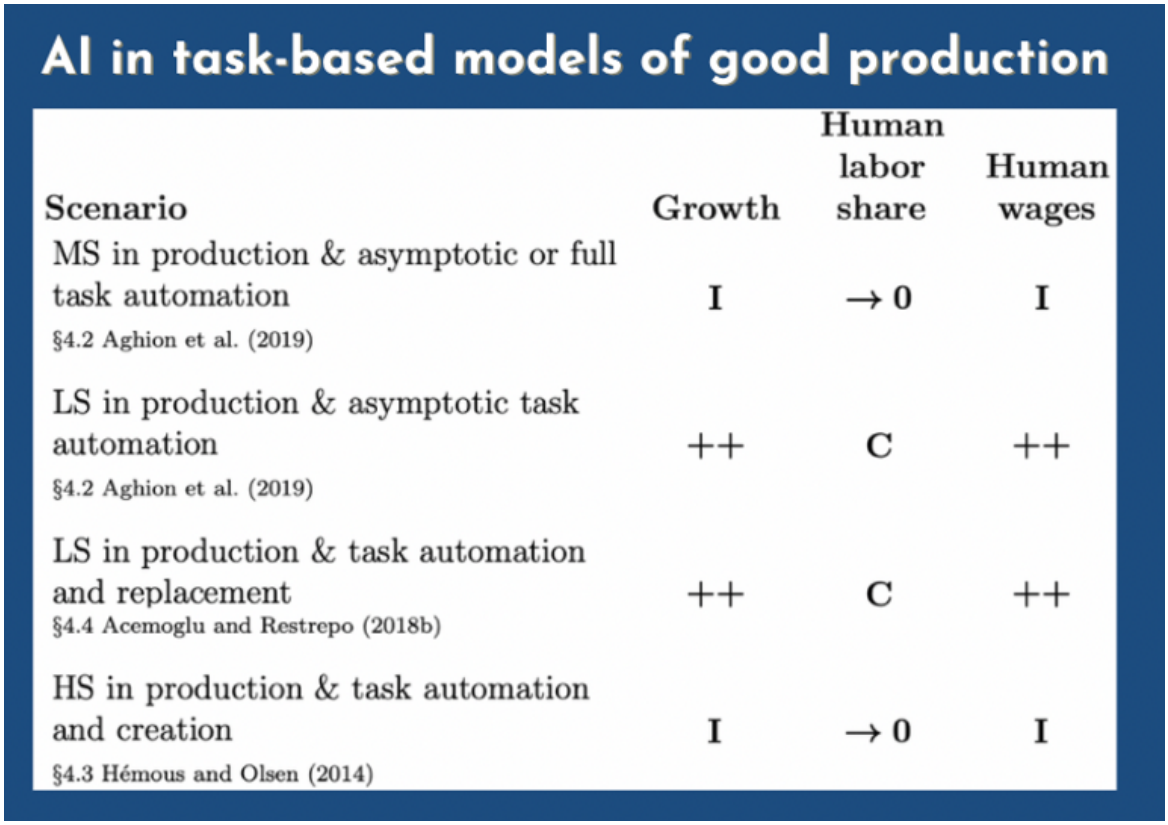
Scenario	Growth	Human labor share	Human wages
PS in production & one-off capital-augmenting tech increase – saving increase §3.5 Korinek and Stiglitz (2019)	++	$\to 0$	=
PS in production & capital-aug. tech growth $\to$ saving decrease §3.5 Sachs and Kotlikoff (2012), Sachs et al. (2015)	–	$\to 0$	$\to 0$

Michael: Then there's the the impacts on savings, oh sorry, how savings can have an impact on growth as well

Phil: so this is also a little bit tangential but maybe worth noting briefly. So in a world where capital can substitute for labor well enough, so if rho is greater than zero, the savings rate can affect the growth rate. Remember before for labor and capital, where gross complements... where rho is negative basically, what drove growth in output per capita, was growth in B. And if the saving rate is too low, then growth is bottlenecked, but as long as the saving rate is enough to keep up with growth in B, then further saving doesn't increase the growth rate. But, once rho is greater than zero, then you can just pile up capital and get more and more output. So increasing the savings rate does increase the growth rate, and it could be that something about AI causes people to save more, or save less than they were saving before. And there's a lot of stories you could tell as to how that would happen. And there are two of them that people have written papers about, but in principle you can imagine a thousand ways that AI would affect the savings rate and it would then affect the growth rate, if rho is greater than zero and...

Michael: when we mean saving, it's not people literally putting money on their bank account or... it's more people investing in S&P 500 or companies right?

Phil: that's right. It has to be am investment. People use the term saving rate, but more accurately would be investment. To say it would be invested with...



Michael: and then there's this whole, I think, this whole literature review part on task-based models for good productions. Could you explain briefly what that is?

Phil: Sure, so as we mentioned a while back, the growth increase... the growth rate increase that we've been benefiting from since the industrial revolution I think, can be pretty well thought of as capital doing ever more tasks that labor used to have to do. So the steam engine or whatever, we now can make machines that do things that people used to do, and at first glance that sounds a very different model of what's been going on with growth, than the model we were talking about before, in which you have labor augmenting technology and capital augmenting technology, and, there's no tasks in that model. You just have the

factory with two pipes in one pipe out, and remember capital augmenting technology was meant... you could use less steel for the cars, maybe labor augmenting technologies, one person figures out a technique to operate two machines at once, but where's the substitute? Where's the task replacement? It's not clear. But, it turns out, a paper from just two years ago shows, that the two are basically equivalent. So if rho is negative (if capital and labor are gross complements) then you can basically... if you imagine that what's going on inside the factory is a whole sequence of tasks, some of which you can use labor, sorry, some of which you can use capital for, and the rest of which you have to use labor for, then as tasks get automated, that ends up functioning like labor augmenting technology. And the way it works is I think pretty interesting. So let's say we start out with 40% of the tasks being done by capital, and the other 60% being done by labor. Because you've only automated the first 40% so far. Then, on average each person sort of has to spread their work across six... this is a hundred tasks... they spread it across sixty tasks

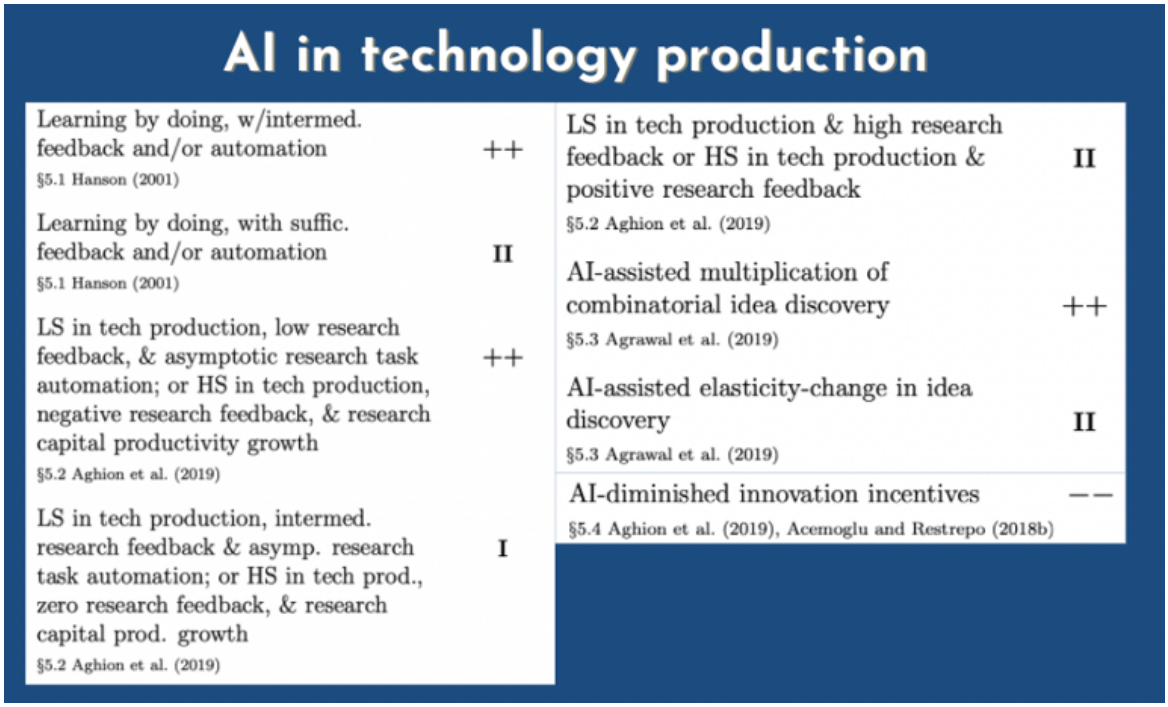
Michael: because they're not automated.

Phil: they're not automated right, and so for every person you add into the person pipe of the big factory of the economy, output only increases by a little bit, because that person has to spread their work across a lot of tasks. In practice they would specialize, but on average you understand each person in effect is spreading their labor across a lot of tasks and you just have a little bit of capital on the side doing the other 40% tasks, and output only goes up a little. But then let's say you automate half of those 60 tasks, so now only 30 tasks are done by labor. Now each person in effect does it... can be responsible for twice as much output as before, because they only need to spread their work across half as many tasks. Capital takes care of the rest, so it's sort of doubling B. You see that?

Michael: I get the intuition.

Phil: and you can have this exponential growth in B, that we've observed, just by having this constant fall in the number of remaining non-automated tasks. So that goes to zero. The number of tasks you need labor for goes to zero asymptotically, but you never quite get there. You always need people for some little bit at the end. And that ends up kind of being the same as growth in B. So anyway, you have all these task-based models of AI, as they're called. And if you just introduce a task-based model to a model with growth in B, separately. So you have automation, and people getting more productive at each task, then that increases the growth rate, and it shouldn't change anything about the labor share. So what does that mean for wages? Well, if the growth rate rises, and the fraction going to labor stays the same, then wages have to rise. So those are those two scenarios in the middle there. But if you kind of automate everything, if you automate all the tasks, or if you kind of shake things up more deeply, then you can get a singularity, rather than just more growth

Michael: and that's when growth and wages grows super exponentially, and we have seen a singularity of type I. I think I got... I think it's interesting that we can make a parallel with task based models and the other ones.



Michael: I haven't even looked too much into AI in technology production.

Phil: well, I think this is kind of the most interesting part of it.

Michael: Okay so go for it.

Phil: we get to self-improving technology. As we can see there are a lot of different models here, but here's where we're finally seeing some infinite output in finite time. And what's going on there is, remember, these semi-endogenous models from before where, you need growth in the number of scientists to maintain output. Well, if you have robot scientists, then basically you can get this feedback loop where you have super exponential growth in scientists, leading to super exponential growth in technology. But if the technology is capital augmenting, and the scientists are made of capital, that means effectively you have more scientists next year. And that's the real crazy feedback loop that could lead to an AI singularity.

Michael: And I believe people like to talk about scientists as something very abstract that is hard to automate because the scientist walks into a bar, has a wife, and does a bunch of human things, but if you just replace science in AI with people coding things, and slapping their fingers at computers, and getting more feedback from what they do... I feel those those robot scientists could just be AI generating code... AI generating AI code is already for me a robot scientist in some sense.

Phil: sure. We need more than computers at the moment to actually develop products and stuff, and have a better understanding of the world so... I think you're frozen

Michael: oh sorry

Phil: no you're there.

Michael: Yeah so it depends on on exactly where what we're focusing on. If we're talking about creating products that can be consumed, yes we need humans to do the carrying, the cooking, and all of those things. But if we're talking about production of software, the AI can... the AI and the human are in a loop where the AI now can suggest code for the human

to read and the human then understands it and accept or not. The code degenerated and... I don't see much more than this, two things human accepting or not accepting and the AI suggesting more code.

Phil: But that's if we're just making code.

Michael: But you don't need much more than code to create an AI, it needs the servers and electronics and net networking behind it.

Phil: sure but that's all for AI getting better and better at being AI. What I'm saying is one thing that you need to have an explosion in scientists more broadly speaking is you need systems, I keep calling them robots but whatever, you need systems that can do all of the things we call AI R&D, not only some of AI R&D. We need... oh we ultimately need a system that can replace human contributions to developing new kinds of computer ships and... whatever

Michael: I think the control argument to that is as what we're actually interested is having AI that can automate human intelligence, and so as long as we have something as general as a human and can think and communicate with humans, and strategize, and those things, then I believe that if we only have code that that self-replicates and becomes better at coding until it reaches just human level intelligence, at this point you can just ask humans to do stuff for him, or manipulate humans to do stuff for him, and so you don't really need more food or human science because now the AI can do whatever he wants by just sending requests or orders to other parts of the world.

Phil: But if there's only so many people, and if people actually remain necessary, then

Michael: oh it wouldn't be enough... it wouldn't be enough to sustain an exponential growth. I got you.

Phil: I think it's worth remembering that intelligence isn't the only input to at least... unless you define intelligence more broadly than where we naturally do intelligence is the only input to R&D... when coming up with new products for instance. Having an intimate understanding of human preferences strikes me as an important, an important factor that machines probably won't have for a long time, or maybe a short time, but it'll be one of the later things for it to beat us up. You might have a friend who's very smart, smarter than you, but doesn't know your preferences as well as you do, and likewise I think we could have an AI that's sort of more intelligent than any human being in every sort of every dimension we call intelligence, but doesn't really have quite as sophisticated... an intuitive understanding of what product will really tickle our fancy, and you would need that to kind of sell this product to humans... just to think of the product in the first place. Okay, oh this is going to be a really good one, oh that's not gonna, no one's gonna use that okay, I'm just saying to really replace people for everything you don't... we casually use the term artificial intelligence but we're not just about intelligence, and you would need artificial arms as well, and we're familiar with that problem. The difficulty of building robots with as much dexterity as people have, that's not the same as intelligent and, and just another thing is there's this intimate understanding of human preferences which... intelligence can help you get that, but it's not identical to intelligence so...

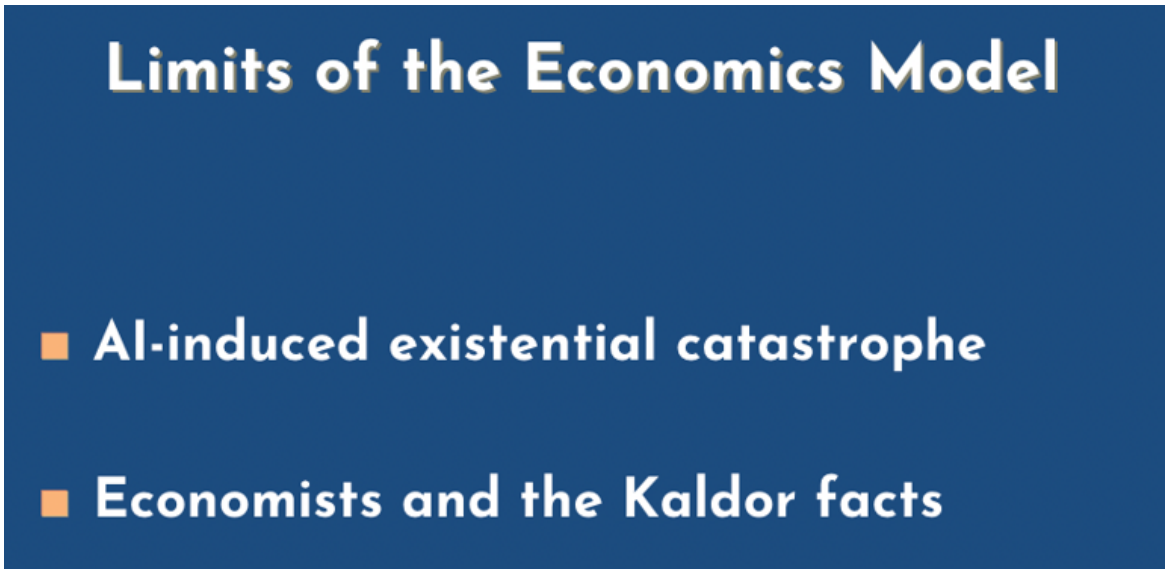
Michael: I think there's this definition of intelligence as more closer to what humans call IQ and ability to think in abstract and logical ways, and then there's this more general definition from, I think, used by Bostrom in his book Superintelligence, as the ability to achieve one's goals, where if you don't have robotic arms... if you don't have human arms, and not with as much dexterity, and you're not able to achieve one's goals, then you're considered less smart than someone else. But in a common sense that doesn't count as intelligence.

Phil: I think that's right we didn't think of Stephen Hawking as less intelligent because he was in a wheelchair... if anything that made him seem more intelligent. Also, the ability to achieve your goals, I think that's like, even if a machine had all the things we conventionally

call intelligence and had arms and was much better than a person achieving their goals, whether people will use it to replace all their human scientists and developers depends on whether it's better than a human at achieving the robot owner's goals and the robot owner's goals might be to come up with better better potato chips, and the robot might not know

Michael: okay... so I guess it doesn't doesn't really... you can have something that is intelligent or or able to stuff in the abstract but you need to have humans that are actually interested in using the thing, for other humans products or preferences that's interesting

Phil: that's right, I think that's sometimes overlooked and that's not relevant if you're Bostrom and you're writing about Superintelligence, you're not even thinking about a scenario in which people continue to own and fully control what these things do, and just decide to use them for this and not for that, we're thinking about the danger that comes with the robot that can have scary goals, and execute on them and... for that all you care about is that it's smart and it has arms. You don't care about whether it knows what kind of potato chips people like... but if we're talking about what effect it will have on the growth rate in a scenario where it stays under control, then you have to think about this other thing as well so...



Michael: I think that that's a perfect transition to our the conclusion of of this whole presentation where, I think at the end of your little literature review you mention that, or maybe it's in the "here this idea" podcast, that economists don't really consider long-term growth increases, and they mostly consider that it will maybe increase by a few percent, but they don't consider dramatic increases, lead apart singularities or something. So, yeah, do you get the impression that there's more and more paper about that, and economists are starting to consider those scenarios? Or are you the only one that... okay not only one because you've done a literature review, but are there more papers being written about this?

Phil: I don't think so. I kind of hoped so. When I wrote this thing I saw a number of papers that considered these singularitarian scenarios but... I was actually at a little online conference... not a pretty big, but online conference on the econ of AI last week hosted by the NBER, National Bureau of Economic Research, and, Anton Kornick, the co-author for the literature review, did a survey of all of us there, on questions about like, "what we thought would happen to the growth rate given AGI" and so on, and only a small number of people thought there was any more than a one percent chance, let alone a 10% chance that AGI would lead to any kind of singularity, and people's projections about the growth rate over the next century (or I think next century was what we were asked about) were very boring. People thought it was almost certainly going to stay around two percent a year.

Michael: what was the population what were there singularitarians or did you have more traditional economists?

Phil: it was entirely traditional economists.

Michael: and what do you think? Are you part of the one percent? Or do you have a more nuanced opinion?

Phil: No, yeah, I think conditional on AGI, the chance of a singularity is more than 10%. I don't think it's like... I don't think it's guaranteed. I forget what number I put down but I've sort of wavered on this, but maybe I think there's a one in four chance of it or something.

Michael: Okay it's 25 percent, yeah it's a bit hand-wavy. We don't have any quantitative estimates on that. I guess it's more kind of more a gut feeling. One thing that makes me laugh at your conclusion in your literature review is that you mentioned that, as long as humans are still a little bit useful for work, they could just stop working and go on strike and then receive a lot of human labor, oh sorry a lot of capital... or wages, and we could have, yeah, things where you can just invest more capital and get like self-improving robots... do you think like any of those two scenarios might happen in your lifetime? Humans going on strike or like, self-improving robots?

Phil: so humans going on strike, certainly.

Michael: Against robots?

Phil: I don't know, if you imagine an amazon packing facility in which the workers aren't being overseen by a human manager, but by a computer, or by a a camera that has some AI ability to see how hard they're working, I can totally see the workers going on strike in that context, and are they going on strike against the robot? Well, sort of. The way we'd think of it is that they're going on strike against the owner of the robot. The owner of the camera, but unless the robots have like completely taken over, it's sort of the same thing, and even once even if robots do take over why would someone be less willing to go on strike because jeff bezos owns the camera than if the camera owns itself...

(If you enjoy those kind of technical videos with diagrams let me know in the comments so I make more of them.)