

# Kickstarting AI

## Technical Journal Entry Begins

### Setting the Vision

It's time to push forward with my work, taking it to the next level. I'm going into the weekend and I have an opportunity to really bear down on the work in that flow-state. Once I get into it, it will take on its own momentum, keeping me engaged. But I have to get the vision down to provide the spark.

I feel the need to do one thing well, but with the pace of change it seems like things are being pulled in many directions. New tools, new skills, constant recalibration. This works against growing expertise and skills that pay you back for years on end. I want to boil things down, but I cannot boil the ocean.

### The Twenty Year Plan

Yep, that pretty much frames it. I need to zero in on a set of skills that have twenty years or so in it. I'm in my mid-50s and if I bear down on my work beginning now in a way that has compounding returns and keep at it to my mid-70s, I will have caught up with not having been on an optimized path earlier in my life. It's a big catch up during a time period when things really change.

I don't want it to be sour grapes, but it's possible that having not been on an optimized path earlier in life could be an advantage today. It means I am still hungry. It means I have all the skills and know-how accumulated up internally. I have invested in myself and have all the right skills for the direction the world is turning.

### Tools and Dependencies

I don't want to have to keep recalibrating on newer and newer tools. I particularly don't want to be at the mercy of this particular vendor or that, trying to bend the world to their will. I do want to use fewer tools better and apply them across a broader set of problems. It's the Unix way. Fewer commands that can be used in more ways and chained together. But I have to remain flexible and take up wacky new tools and interfaces here and there like the shiny new toy crowd who drop the current thing and run to the next as a sort of sport.

### The Reality of Platform Dependencies

It's painful, but you have to use different systems and not all of them free of annoying dependencies. The question is how much you invest your muscle memory into the ones that live with you for life versus those with a short lifespan, say 2 to 5 years. Anything Windows and Mac is of that later variety, they need to get you to buy again. The need for recurring revenue means required dependencies, and skills will change, and you will be on a hamster wheel by design. You can't not hop on certain hamster wheels. You've gotta be on them in order to be in the

overall race. Even just to be aware of what others are going through on the hamster wheel, it's worth being on the hamster wheel, a little bit yourself. No matter how fervently free and open source advocate you are, you probably should have one Windows machine, and one Mac machine in your life. It sucks, but that's where the broadest perspective is gonna come from.

## Two Types of Muscle Memory

You have to come to grips with the fact that there are more than one kind of muscle memory, and you need both. There's the kind that makes some tool and extension of your body for your entire life. And there's the kind that is sort of an ad hoc appendage for just as long as you need it, then gets tossed on the trash heap when you're done along with your invested muscle memory. The number of concrete examples I could go of here is endless. The memory skills you get for using the Windows subsystem for Linux is disposable. But the skills you get in the actual Linux terminal are timeless.

## API Wrapping and Value Creation

There is an API wrapping store here. Fundamental tools that have a really long life tend to be fairly difficult because they need to support many edge cases. The urllib library that is built into the Python standard library for making http web requests would be a good example. It has to support all cases, so it is at the expensive easy every day use "for humans", so few people use it directly. The massively popular Requests library comes along, and everyone hops on that bandwagon because it has a better API. It "wrapped" urllib's API in one that's make "for humans" and won users.

Such API wrapping has various purposes. The obvious one is making the underlying tool easier than more accessible to more people for the common use cases. As you might imagine, this is also an opportunity to extract money from users who would not otherwise have to pay for the underlying tool. So the API wrapping layer in those cases is commercial, proprietary and intended to extract money. It's a business.

## The AI Revolution

Probably the biggest example today are things that wrap AI APIs and provide a convenient user interface layer, like the Cursor editor. Well, I'm not downplay the amount of work or innovation that can go into API wrapping, it's still essentially wrap-and-charge. When you examine the value chain, the progression of all materials to things of greater value, the most profitable parts are where the least effort and cost is required to take the wrong materials, and the most money can be charged for some processing of that raw materials into more valuable finished product.

## The Raw Materials of AI

The raw material in AI, the new Industrial Revolution, is compute at scale — hyper-scalers, they're called. The investment here is enormous. And even when you are the world's traditional experts like Intel, you still might lose because you've got some little detail wrong or failed to prepare for this new world with a CUDA library winning the hearts and minds of developers for decades. In other words, mining the raw material of the age of AI is a big boy game. Nvidia basically won. Tesla, OpenAI, Google, Meta and just about everyone else who can pull off such

hyper-scaling projects are looking at how to break this Nvidia dependency. And they will, but it's gonna take ages to overcome the early move advantage and big hardware/software/habits trilemma moat advantage.

## **Local AI and Distributed Computing**

Some of the AI revolution will be fought with distributed computing in people's homes on ragtag computers of infinite variety. They will be a heterogeneous mix of AMD, Groq, and whatever else weird hardware can run an inference engine. That is to say, local AI. The models, that is to say intelligent entities, "baked" on the hyper scaler hardware can basically run anywhere, and anyone's home as a Genie In a Bottle. You don't need to be a big boy to play with big boy toys.

## **The Future of Data Centers**

And that brings us to this morning's ponderings. I'm watching Satya Nadella's interview on Dwarkesh Patel's podcast on YouTube. He lays out the future very well. It's clear that lots of data centers have to be built in lots of places around the world, and those data centers have to be used to train and test models, and then repurposed for inference, and somehow do it at a profit. Folks can't figure out how to utilize this infrastructure to add more value to things in the value chain, then it's a giant bust. Of course he didn't say that. But that's the fact of it. It's going to be exactly like Amazon overbuilding capacity after an embarrassing holiday season, and then reselling that capacity when it didn't need it as AWS. Same thing. But AI.

## **The Hardware Arms Race**

There is a rolling Moore's Law effect here too. If there is a competitive advantage to be had by utilizing such hyper scaling resources, then there is more competitive advantage to be had by making sure the data center hardware is keeping pace with Moore's Law. And all the depreciated and discarded hardware will go to the second tier bottom scraper hosts, if they don't just recycle it to keep it out of competitors hands. But in either case, the hardware data center arms race isn't gonna stop, even when AI computer is completely commoditized. The big boy game becomes even more big boy, and the fusion generators and such are going to have to be employed to keep it all churning.

## **The Energy Question**

Sci-Fi? Well, yeah of course. Isn't it all at this point? France just got a fusion reaction up to 20 minutes with their tokamak fusion reactor. The amount of investment going into all kinds of traditional and wacky energy is off the charts. If there is a value to be extracted by continuous upscaling of hyper scaling then the positive feedback loop is at the beginning of its hockey stick curve. I will remind you that the one over X power law curve that makes you think you missed out of investment opportunity always looks like one over X no matter how far in or far out you scale. We are right at the beginning. This is the world we need to prepare for.

## **Wealth Distribution and Human Nature**

Of course, there is a damping effect. The damping effect being primarily human nature and

resource distribution. The infinitely have's of the one percent become even more infinitely have's, while the infinitely have-not's become even more destitute in their have-nots-ness. That's just human nature. That's just wealth concentration power law dynamics. It's Pareto distribution, and would've been the natural state of things in the world where not for FDR and the new deal in the G.I. bill that red distributed wealth, especially to soldiers coming home from World War II. The suburbs and new expectations were artificially architected. The pendulum swings both ways, and given the removal of the artificial force compelling the middle class, it's swinging back to a more natural state.

## **The Nature of Resource Hoarding**

Of course, this sounds ferociously evil. It sounds dystopian. But the truth is those who obsess over money and hoarding resources and get down the trick of passing it down generationally to their children, can have a massive snowballing effect. What became important to the parents becomes important to the child. All kinds of legacy minds tricks gets the children perpetuating this momentum. It becomes self-sustaining. Is this evil? Or is this just basic tribal dynamics on a massive scale over generations? Is this just basic primate dynamics? Chimpanzees? Anyone watching chimps in those documentaries might come to a conclusion that they're evil. Isn't it just naturally selected behavior? Perhaps, gone unchecked.

One could even go as far as to say the tribe not keeping their powerful resource hoarders in check is the actual evil. The failure to occasionally re-level the playing field. Unfortunately, that's done through war. Or through communism, which inevitably leads to corruption and totalitarianism. So you're caught between a rock and a hard place, and the rich get richer. All roads lead to corruption. All roads lead to radically inequitable distributions of wealth. Why do they haves hoarde and the have nots not? Is that an impossible to change aspect of human nature without war or tyranny?

## **The Evolution of AI Thought**

We are seeing human thought get reconstructed. One-shot prompt/responses are the equivalent of baby-babble. As coherent as it might seem, it's the result of a baby-like intelligence having access to a vast store of data, so that it can formulate its responses to sound coherent through imitation. But reflection and chain of thought reasoning of the second round, starting on the publicly witnessable side with OpenAI ChatGPT o1 model, brings it from baby babble to adult rational thought. It's what struck so deep with DeepSeek, taking this proprietary seeming reflection process o1 does and making it lightweight enough to run on your Macbooks and Windows laptops.

Another way of looking at it is that it's not the massive depth and breath of the hundreds-of-gigabyte trained models. It's the lightweight inference engines doing the best they can with back-and-forth self-reflection! Sure, a big beefy data-laden starting point helps. But for your money, it's also important to take the output of a system and recursively feed it back to the input of the same system, letting it bake on a task with a good set of rules until it spirals inwards to a self-satisfying answer. And the self that is satisfied is the inference engine itself!

## **The State of AI Censorship**

Google Gemini is highly censored. Even the more advanced models can't name the president of

the United States if asked. And it's not because of the training cutoff date. It's got specific keyword filtering in place. It's utterly ridiculous, a month into the new presidency. It's eroding Gemini's credibility in my mind, no matter how smart it is. More specifically, it's eroding Google's credibility in my mind. Don't be evil? Evil thus announces itself,ahaha! They are the epitome of "if you don't pay for the product, you are the product." And this is particularly sad and poignant given how vocal about and opposed to that very issue they were in their early days. And then they went public.

## **Planning for the Future**

There's so much to unpack there, but a critique of capitalism isn't my purpose with this article. It's to plan and plot my course in the wake of capitalism's recovery. Yup. I'm sold. There is a recovery going on, and I didn't realize how bad it was. My struggles with Gemini's prompt rejections gloriously reinforces a narrative I didn't want to believe. The suppression of the details of a certain laptop by certain organizations at certain critical moments, sold as enemy state propaganda but which wasn't... after Musk bought Twitter and looked at what was suppressed and Zuckerberg spoke up to say Facebook faced the same heat, and how advertisers boycotted Twitter... well, even the most "I don't want to believe" hold-outs have to let rational thought win eventually. A correction is occurring, and that's part of what we have to factor into the next moves that push forward my work.

## **Historical Context and Wealth Distribution**

World War II is long since over. The 40 years of 2 generations coming and going, long enough to forget, has long since past. In fact if you measure the end of WWII from 1945, then it's been almost double that. 80 years is twice the everything-is-forgotten barrier and is approaching the "no one is alive anymore who might remember" line. The former is a barrier or an obstacle, because plenty of people with firsthand knowledge are still around who can give you better lessons than the history books. The second is a line, because they're gone. What was recorded was recorded and what wasn't is lost, except to anthropologists who can dust off the occasional unintentional or undiscovered recordings.

The significance of WWII so far behind us, and really the Great Depression of 1929 and the 30s (we're nearly in the 30s again, 100 years later!) is that Franklin D. Roosevelt's New Deal that birthed middle-class America is long behind us as well. The significance of that is that the aforementioned natural shape of wealth distribution according to human nature, the power-law, the Pareto Principle of Vilfredo Pareto, or however else you want to look at it suggests the have's will accumulate, the have-not's will lose, and those in the middle will stratify towards becoming a have or a have-not. This is just basic primate tribal dynamics per chimpanzees and bonobos in place here. If it's not violence being used to keep a social hierarchy well defined, then it's sex.

## **Language and Intelligence**

Yup. We're not that different from our relatives, but we have a veneer of rational thought superimposed, mostly arising from language. Language is a matter of greater articulation of ideas, expressing them, capturing them, convincing others of them, and so on.

## The Evolution of AI Intelligence

As I watch AI evolve in the public eye, which is a sheer pleasure and very different from most sci-fi dystopian scenarios that fuel our nightmares, I see something very similar. One-shot prompt/replies are baby-talk babble without the articulateness. It's a level of intelligence you can hold all in your mind at once and make pretty good output with. It's convincing, but it's young. It's that level of 4 year-old to 7 year-old intelligence we're always attributing to elephants, parrots, dolphins, octopuses and the like. You can look at something. You can think. You can even introspect a little, having a sense of self (knowing when you're looking at yourself in a mirror), and engage in a bit of that meta-cognition (thinking about thinking) that we so often like to attribute exclusively to humans, but like Jane Goodall blew away the notion that humans were the only tool-makers, has been long blown away as well. Many things including animals and machine intelligences meta-cognate.

## The Current State of AI Models

So, we are in the generation that gets to see the kickstarting of AI. We prompt. Some deterministic pre-trained model responds. Given the exact same prompt and no hanky panky to deliberately randomize output, the same output will be generated for the same input. Today's popular chat-based AI models are these static snapshots. They are files. You'll hear the term "weights". Yeah, even model that you hear named, like ChatGPT 4.5, Gemini 2.0, Grok 3 or whatever is basically a file on a storage device, like an Excel file. They get momentarily **instantiated**, aka loaded up into memory as a running instance of an intelligent, thinking entity, get fed your prompt as input, generates output, and then (generally) gets destroyed. It gets dumped from memory. It's like the way the Web was when it was born, with a technique called cgi-bin, which loaded everything up into memory, did its thing, and then dumped it all. Today's cloud-based AI models productized for the public are a lot like the Mr. Meseeks character from Rick and Morty.

## The Illusion of Persistence

And we are in the generation getting to know these momentarily instantiated static models posing as continuous intelligences, but which aren't. That's not to say that they're not intelligent, nor that they are not entities. They are indeed intelligent, and they are indeed entities. They're just entities that blip in and out of existence as fast as you prompt them. They are ephemeral intelligent entities, blinking in and out of existence as needed at the moment. This is for so many reasons right now, at this state of human history and technology, not the least of which is resource allocation. Like in the early days of the Web and the cgi-bin method of serving websites, things might get so complicated with memory (leaks and all) that the only way to be sure is to nuke the place from orbit. It's the only way to be sure. A fresh state every time. The same "starting point entity" every time.

The main way the illusion of persistence is achieved with today's chatbots is the entire discussion history up to this point (up to the size of the token window size) fed back into it invisibly in the background. In other words, when you think you sent a chatbot just the latest message, you actually **re-posted the entire discussion**. And that's how the illusion of it getting to know you, having any sense of continuity or personalization is achieved... right now, today. And only in most cases. Of course there's edge case exceptions even today, to make a better

product. The exceptions generally have to do with some database dedicated just to you kept in the cloud that the chatbot can invisibly query against in the background in the course of your discussion to fill in some blanks. It's the early stages of layering in memory systems akin to human memory, be it short-term or long-term. The seminal article from Google, "Attention Is All You Need" has a follow-up paper addressing this memory issue called "Titans: Learning to Memorize at Test Time".

## **The Future of AI Memory**

So, memory will be getting better and layered into AI systems in countless different ways. It's resource-constrained on the cloud, especially with free services, and especially concerning on the privacy front with free cloud services where you are the product. Have no doubt, when it's free all your interactions with the AI are being used to train the model in the background, presumably after some personal data scrubbing that they actually want to do for legal liability reasons, but nonetheless pretty much means the price of free is you training the models. Elon Musk buying Twitter and giving that enormous database to Grok 3 to train on is a wonderful example. It also may be one of the early examples of the underlying model not being so static, and not having a precise cutoff date of some batch-file training process -- at least ostensibly. I don't fully believe that yet, but I am hopeful.

## **The State of Free Software**

Anyhow, so I talked about the main way the static models running on architecturally genocidal cloud architectures are working. I talked about why such snuffing out of intelligent entities as fast as they're created is necessary. I touched on a few of the edge cases, and why. And how you get those edge cases only by paying in one way or another, because almost nothing is really free. I ought to talk about the "almost nothing" edge cases where true altruism really does exist in this world and is at the heart of the "Free" part of the "Free and Open Source Software" (FOSS) movement, but I don't want this article to go excessively in that direction. Suffice to say that you will notice that the Free Software Foundation didn't rename itself to the Free and Open Source Software Foundation. Yes, while open is important, IP-burdened (intellectual property burdened) is still IP-burdened. MySQL, Redis, MongoDB, Confluent, Grafana and more. It's a trend. FOSS it, get 'em hooked, and then get the F out of there. But that's for another time.

## **Moving Forward**

And so, this is about adjusting my work directionally. I'm already at a friggin great starting point. I just have to tweak it a bit here and there to make it look and feel that way for the rest of the world, but starting with the folks I work with, to get it into a sort of "right of first refusal" state. That's not to say it will ever be fully proprietary as I have maintained its FOSS licensing on it as a project that predated this. But it does effect where I put my focus. It's part of the directional thing. Do I become a YouTuber again, or do I keep it behind closed doors doing stupid genie tricks for a limited audience? Hmmmm.

Okay, this is getting framed well. Before it can become stupid genie tricks (in the spirit of David Letterman's stupid pet tricks -- boy, I'm getting old), I have to make it completely well described linear workflows can't escape understanding how they work grab 'em by the scruff of the neck and walk them through the process tricks. In other words, the training material. On a starting

point that works as a beachhead or foothold. And that's this weekend. I need that starting nub on the linear workflow.

## Quality and Business Principles

Things tank in quality when you get to love them. The very quality that makes a thing love-worthy gets it onto the radar of opportunistic profiteers looking to maximize income and squeeze that lemon. You'll see this in popular products all the time. They become famous and well known for some quality, and then they ship off the manufacturing to the lowest bidder in China, and BAM! Quality tanks, and this super-awesome product that was a secret weapon of early adopter insiders, often a sort of product-cult, become disenfranchised and alienated, but the company doesn't care because they're raking in the bucks. On rare occasion the story will be the reverse, like Harley Davidson. You want to be that exception, and it come from applying well known business principles from the likes of Peter Ducker and Edwards Demming. The former being the get-and-keep customers guy, and the later being the quality-management guy. Together, you get things like the Harley Davidson bounce back, when it could have gone the downward spiral direction of so many other love-worthy products.

## The Starting Point

I'm trying to make the starting point of my project love-worthy. It's like the joyful framework idea of Ruby on Rails, but updated for Python in a post-template, post-JavaScript, post-scaling world. Yup. No templates like jina2 or mako, no JavaScript frameworks like React or Vue, and no arbitrary constraints or limitations on single-instance versions because it has to scale in the cloud. There's a lot of anti-patterns here, and the trick is I have to make the sum of the parts something wholly unique, scratching itches a lot of folks out there feel but don't know they have, and even if they did, can't quite reach.

---

## Gemini 2.0 Advanced Flash Brainstormer

In this article, the author outlines a twenty-year plan to focus on developing skills with long-term value in the rapidly changing tech landscape, particularly in AI. They discuss the need to balance platform dependencies, the importance of API wrapping for accessibility, and the evolving nature of AI, from large-scale data centers to local AI and distributed computing. The author also reflects on wealth distribution, the impact of historical events on current economic structures, and the evolution of AI intelligence, while critiquing the current state of AI censorship and the transient nature of AI entities. They conclude by emphasizing the need to create a "love-worthy" starting point for their project, focusing on quality and unique solutions in a post-template, post-JavaScript, and post-scaling world.

### Ideas for Titles and Headlines:

- "The Twenty-Year Tech Vision: Mastering AI and Beyond"
- "Navigating the AI Revolution: A Long-Term Strategy"
- "Beyond the Hamster Wheel: Building Timeless Skills in Tech"
- "API Wrapping and the Future of AI: A Personal Journey"
- "From Baby Babble to Adult Thought: The Evolution of AI"

**My Opinion of the Article:**

The article presents a thoughtful and wide-ranging perspective on the current state and future of technology, particularly AI. The author's blend of personal reflection, technical insights, and socio-economic commentary creates a compelling narrative. The discussion on the transient nature of AI entities and the critique of AI censorship are particularly interesting. The author's long-term vision and emphasis on building sustainable skills resonate, and the article successfully captures the complexities and opportunities of the rapidly evolving tech world.