Dwarkesh Podcast #82 - Gwern Branwen - How an Anonymous Researcher Predicted AI's Trajectory

Published - November 13, 2024

Transcribed by - thepodtranscripts.com

Dwarkesh Patel
Today I’m interviewing Gwern Branwen. Gwern is an anonymous researcher and writer. He’s deeply influenced the people building AGI. He was one of the first people to see LLM scaling coming. If you’ve read his blog, you’ll know he’s one of the most interesting polymathic thinkers alive. We recorded this conversation in person. In order to protect Gwern’s anonymity, we created this avatar. This isn’t his voice. This isn’t his face. But these are his words.

What is the most underrated benefit of anonymity?

Gwern Branwen
The most underrated benefit of anonymity is that people don't project onto you as much. They can't slot you into any particular niche or identity and write you off in advance. They have to at least read you a little bit to even begin to dismiss you.

It's great that people cannot retaliate against you. I have derived a lot of benefit from people not being able to mail heroin to my home and call the police to SWAT me. But I always feel that the biggest benefit is just that you get a hearing at all. You don't get immediately written off by the context.

Dwarkesh Patel
Do you expect companies to be automated top-down (starting with the CEO) or bottom-up (starting with all the workers)?

Gwern Branwen
All of the pressures are to go bottom-up. From existing things, it's just much more palatable in every way to start at the bottom and replace there and work your way up, to eventually where you just have human executives overseeing a firm of AIs.

Also from a RL perspective, if we are in fact better than AIs in some way, it should be in the long-term vision thing. The AI will be too myopic to execute any kind of novel long-term strategy and seize new opportunities.

That would presumably give you this paradigm where you have a human CEO who does the vision thing. And then the AI corporation scurries around doing his bidding. They don't have the taste that the CEO has. You have one Steve Jobs-type at the helm, and then maybe a whole pyramid of AIs out there executing it and bringing him new proposals. He looks at every individual thing and says, “No, that proposal is bad. This one is good.”

That may be hard to quantify, but the human-led firms should, under this view, then outcompete the entirely AI firms, which would keep making myopic choices that just don't quite work out in the long term.

Dwarkesh Patel
What is the last thing you'd be personally doing? What is the last keystroke that gets automated for you?

Gwern Branwen
The last thing that I see myself still doing right before the nanobots start eating me from the bottom up and I start screaming, "No, I specifically requested the opposite of this…." Right before that, I think what I'm still doing is the Steve Jobs-thing of choosing. My AI minions are bringing me wonderful essays. I'm saying, "This one is better. This is the one that I like," and possibly building on that and saying, "That's almost right, but you know what would make it really good? If you pushed it to 11 in this way."

Dwarkesh Patel
If we do have firms that are made up of AIs, what do you expect the unit of selection to be? Will it be individual models? Will it be the firm as a whole? With humans, we have these debates about whether it's kin-level selection, individual-level selection, or gene-level selection. What will it be for the AIs?

Gwern Branwen
Once you can replicate individual models perfectly, the unit of selection can move way up and you can do much larger groups and packages of minds. That would be an obvious place to start. You can train individual minds in a differentiable fashion, but then you can't really train the interaction between them. You will have groups of models or minds of people who just work together really well in a global sense, even if you can't attribute it to any particular aspect of their interactions. There are some places you go and people just work well together. There's nothing specific about it, but for whatever reason they all just click in just the right way.

That seems like the most obvious unit of selection. You would have packages—I guess possibly department units—where you have a programmer and a manager type, then you have maybe a secretary type, maybe a financial type, a legal type. This is the default package where you just copy everywhere you need a new unit. At this level, you can start evolving them and making random variations to each and then keep the one that performs best.

Dwarkesh Patel
By when could one have foreseen the Singularity? Obviously, Moravec and others are talking about it in the eighties and nineties. You could have done it decades earlier. When was the earliest you could have seen where things were headed?

Gwern Branwen
If you want to trace the genealogy there, you'd have to at least go back as far as Samuel Butler's Erewhon in 1872 or his essay before that. In 1863, he describes explicitly his vision of a machine life becoming ever more developed until eventually it's autonomous. At which point, that's a threat to the human race. This is why he concluded, "war to the death should be instantly proclaimed against them." That's prescient for 1863! I'm not sure that anyone has given a clear Singularity scenario earlier than that. The idea of technological progress was still relatively new at that point.

I love the example of Isaac Newton looking at the rates of progress in Newton's time and going, "Wow, there's something strange here. Stuff is being invented now. We're making progress. How is that possible?" And then coming up with the answer, "Well, progress is possible now because civilization gets destroyed every couple of thousand years, and all we're doing is we're rediscovering the old stuff."

That's Newton's explanation for technological acceleration. We can't actually have any kind of real technological acceleration. It must be because the world gets destroyed periodically and we just can't see past the last reset.

Dwarkesh Patel
It's almost like Fermi's paradox, but for different civilizations across time with respect to each other instead of aliens across space.

Gwern Branwen
Yeah. It turns out even Lucretius, around 1,700 years before that, is writing the same argument. "Look at all these wonderful innovations and arts and sciences that we Romans have compiled together in the Roman empire! This is amazing, but it can't actually be a recent acceleration in technology. Could that be real? No, that's crazy. Obviously, the world was recently destroyed."

Dwarkesh Patel
Interesting.

Gwern Branwen
It is, it is.

Dwarkesh Patel
What is the grand parsimonious theory of intelligence going to look like? It seems like you have all of these trends across different fields—like scaling laws in AI, like the scaling of the human brain when we went from primates to humans, the uniformity of the neocortex—and basically many other things which seem to be pointing towards some grand theory that should exist which explains what intelligence is. What do you think that will look like?

Gwern Branwen
The 10,000 foot view of intelligence, that I think the success of scaling points to, is that all intelligence is is search over Turing machines. Anything that happens can be described by Turing machines of various lengths. All we are doing when we are doing "learning," or when we are doing "scaling," is that we're searching over more and longer Turing machines, and we are applying them in each specific case.

Otherwise, there is no general master algorithm. There is no special intelligence fluid. It's just a tremendous number of special cases that we learn and we encode into our brains.

Dwarkesh Patel
I don't know. When I look at the ways in which my smart friends are smart, it just feels more like a general horsepower kind of thing. They've just got more juice. That seems more compatible with this master algorithm perspective rather than this Turing machine perspective. It doesn't really feel like they've got this long tail of Turing machines that they've learned. How does this picture account for variation in human intelligence?

Gwern Branwen
Well, yeah. When we talk about more or less intelligence, it's just that they have more compute in order to do search over more Turing machines for longer. I don't think there's anything else other than that. So from any learned brain you could extract small solutions to specific problems, because all the large brain is doing with the compute is finding it. That's why you never find any "IQ gland". There is nowhere in the brain where, if you hit it, you eliminate fluid intelligence. This doesn't exist. Because what your brain is doing is a lot of learning of individual specialized problems. Once those individual problems are learned, then they get recombined for fluid intelligence. And that's just, you know… intelligence. Typically with a large neural network model, you can always pull out a small model which does a specific task equally well. Because that's all the large model is. It's just a gigantic ensemble of small models tailored to the ever-escalating number of tiny problems you have been feeding them.

Dwarkesh Patel
If intelligence is just search over Turing machines—and of course intelligence is tremendously valuable and useful—doesn't that make it more surprising that intelligence took this long to evolve in humans?

Gwern Branwen
Not really, I would actually say that it helps explain why human-level intelligence is not such a great idea and so rare to evolve. Because any small Turing machine could always be encoded more directly by your genes, with sufficient evolution. You have these organisms where their entire neural network is just hard-coded by the genes. So if you could do that, obviously that's way better than some sort of colossally expensive, unreliable, glitchy

search process—like what humans implement—which takes whole days, in some cases, to learn. Whereas you could be hardwired right from birth.

For many creatures, it just doesn't pay to be intelligent because that's not actually adaptive. There are better ways to solve the problem than a general purpose intelligence.

In any kind of niche where it's static, or where intelligence will be super expensive, or where you don't have much time because you're a short-lived organism, it's going to be hard to evolve a general purpose learning mechanism when you could instead evolve one that's tailored to the specific problem that you encounter.

Dwarkesh Patel
You're one of the only people outside OpenAI in 2020 who had a picture of the way in which AI was progressing and had a very detailed theory, an empirical theory of scaling in particular. I'm curious what processes you were using at the time which allowed you to see the picture you painted in the "Scaling Hypothesis" post that you wrote at the time.

Gwern Branwen
If I had to give an intellectual history of that for me, it would start in the mid-2000s when I'm reading Moravec and Ray Kurzweil. At the time, they're making this kind of fundamental connectionist argument that if you had enough computing power, that could result in discovering the neural network architecture that matches the human brain. And that until that happens, until that amount of computing power is available, AI is basically futile.

To me, I found this argument very unlikely, because it's very much a "build it and they will come" view of progress, which at the time I just did not think was correct. I thought it was ludicrous to suggest that simply because there's some supercomputer out there which matches the human brain, then that would just summon out of nonexistence the correct algorithm.

Algorithms are really complex and hard! They require deep insight—or at least I thought they did. It seemed like really difficult mathematics. You can't just buy a bunch of computers and expect to get this advanced AI out of it! It just seemed like magical thinking.

So I knew the argument, but I was super skeptical. I didn't pay too much attention, but Shane Legg and some others were very big on this in the years following. And as part of my interest in transhumanism and LessWrong and AI risk, I was paying close attention to Legg's blog posts where he's extrapolating out the trend with updated numbers from Kurzweil and Moravec. And he's giving very precise predictions about how we're going to get the first generalist system around 2019, as Moore's law keeps going. And then around 2025, we'll get the first human-ish agents with generalist capabilities. Then by 2030, we should have AGI.

Along the way, DanNet and AlexNet came out. When those came out I was like, “Wow, that's a very impressive success story of connectionism. But is it just an isolated success story? Or is this what Kurzweil and Moravec and Legg were predicting— that we would get GPUs and then better algorithms would just show up?”

So I started thinking to myself that this is something to keep an eye on. Maybe this is not quite as stupid an idea as I had originally thought. I just keep reading deep learning literature and noticing again and again that the dataset size keeps getting bigger. The models keep getting bigger. The GPUs slowly crept up from one GPU—the cheapest consumer GPU—to two, and then they were eventually training on eight.

And you can just see the fact that the neural networks keep expanding from these incredibly niche use cases that do next to nothing. The use just kept getting broader and broader and broader. I would say to myself, “Wow, is there anything CNNs can't do?” I would just see people apply CNN to something else every individual day on arXiv.

So for me it was this gradual trickle of drops hitting me in the background as I was going along with my life. Every few days, another drop would fall. I’d go, “Huh? Maybe intelligence really is just a lot of compute applied to a lot of data, applied to a lot of parameters. Maybe Moravec and Legg and Kurzweil were right.” I’d just note that, and continue on, thinking to myself, “Huh, if that was true, it would have a lot of implications.”

So there was no real eureka moment there. It was just continually watching this trend that no one else seemed to see, except possibly a handful of people like Ilya Sutskever, or Schmidhuber. I would just pay attention and notice that the world over time looked more like their world than it looked like my world, where algorithms are super important and you need like deep insight to do stuff. Their world just kept happening.

And then GPT-1 comes out and I was like, “Wow, this unsupervised sentiment neuron is just learning on its own. That's pretty amazing.” It was also a very compute-centric view. You just build the Transformer and the intelligence will come.

And then GPT-2 comes out and I had this “holy shit!” moment. You look at the prompting and the summarization: “Holy shit, do we live in their world?

And then GPT-3 comes out and that was the crucial test. It's a big, big scale-up. It's one of the biggest scale-ups in all neural network history. Going from GPT-2 to GPT-3, that's not a super narrow specific task like Go. It really seemed like it was the crucial test. If scaling was bogus, then the GPT-3 paper should just be unimpressive and wouldn't show anything important. Whereas if scaling was true, you would just automatically be guaranteed to get so much more impressive results out of it than GPT-2.

I opened up the first page, maybe the second page, and I saw the few-shot learning chart. And I'm like, “Holy shit, we are living in the scaling world. Legg and Moravec and Kurzweil were right!”

And then I turned to Twitter and everyone else was like, “Oh, you know, this shows that scaling works so badly. Why, it's not even state-of-the-art!” That made me so angry I had to write all this up. Someone was wrong on the Internet.

Dwarkesh Patel
I remember in 2020, people were writing bestselling books about AI. It was definitely a thing people were talking about, but people were not noticing the most salient things in retrospect: LLMs, GPT-3, scaling laws. All these people who are talking about AI but missing this crucial crux, what were they getting wrong?

Gwern Branwen
I think for the most part they were suffering from two issues. First, they had not been paying attention to all of the scaling results before that which were relevant. They had not really appreciated the fact that, for example, AlphaZero was discovered in part by DeepMind doing Bayesian optimization on the hyperparameters and noticing that you could just get rid of more and more of the tree search as you went and you got better models. That was a critical insight, which could only have been gained by having so much compute power that you could afford to train many, many versions and see the difference that that made.

Similarly, those people simply did not know about the Baidu paper on scaling laws in 2017, which showed that the scaling laws just keep going and going forever, practically. It should have been the most important paper of the year, but a lot of people just did not prioritize it. It didn't have any immediate implication, and so it sort of got forgotten. People were too busy discussing Transformers or AlphaZero or something to really notice it.

So that was one issue. Another issue is that they shared the basic error I was making about algorithms being more important than compute. This was, in part, due to a systematic falsification of the actual origins of ideas in the research literature. Papers do not tell you where the ideas come from in a truthful manner. They just tell you a nice sounding story about how it was discovered. They don’t tell you how it’s actually discovered.

So even if you appreciate the role of trial and error and compute power in your own experiment as a researcher, you probably just think, “Oh, I got lucky that way. My experience is unrepresentative. Over in the next lab, there they do things by the power of thought and deep insight.”

Then it turns out that everywhere you go, compute and data, trial and error, and serendipity play enormous roles in how things actually happened. Once you understand that, then you understand why compute comes first. You can't do trial and error and serendipity without it. You can write down all these beautiful ideas, but you just can't test them out.

Even a small difference in hyperparameters, or a small choice of architecture, can make a huge difference to the results. When you only can do a few instances, you would typically find that it doesn't work, and you would give up and you would go away and do something else.

Whereas if you had more compute power, you could keep trying. Eventually, you hit something that works great. Once you have a working solution, you can simplify it and improve it and figure out why it worked and get a nice, robust solution that would work no matter what you did to it. But until then, you're stuck. You're just flailing around in this regime where nothing works.

So you have this horrible experience going through the old deep learning literature and seeing all sorts of contemporary ideas people had back then, which were completely correct. But they didn't have the compute to train what you know would have worked. It’s just tremendously tragic. You can look at things like ResNets being published back in 1988, instead of 2015.

And it would have worked! It did work, but at such a small scale that it was irrelevant. You couldn't use it for anything real. It just got forgotten, so you had to wait until 2015 for ResNets to actually come along and be a revolution in deep learning.

So that’s kind of the double bias of why you would believe that scaling was not going to work. You did not notice the results that were key, in retrospect, like the BigGAN scaling to 300 million images. There are still people today who would tell you with a straight face that GANs cannot scale past millions of images. They just don't know that BigGAN handled 300 million images without a sweat. If you don't know that, well you probably would easily think, “Oh, GANs are broken.” But if you do know that, then you think to yourself, “How can algorithms be so important when all these different generative architectures all work so well—as long as you have lots and lots of GPUs?” That's the common ingredient. You have to have lots and lots of GPUs.

Dwarkesh Patel
What do your timelines look like over the last 20 years? Is AI just monotonically getting closer over time?

Gwern Branwen
I would say it was very far away, from like 2005 to 2010. It was somewhere well past like 2050. It was close enough that I thought I might live to see it, but I was not actually sure if there was any reasonable chance.

But once AlexNet and DanNet came out, then it just kept dropping at a rate of like 2 years per year, every year until now. We just kept on hitting barriers to deep learning and doing better. Regardless of how it was doing it, it was obviously getting way better. It just seemed none of the alternative paradigms were doing well. This one was doing super well.

Dwarkesh Patel
Was there a time that you felt you had updated too far?

Gwern Branwen
Yeah, there were a few times I thought I had overshot. I thought people over-updated on AlphaGo. They went too far on AI hype with AlphaGo. Afterwards, when pushes into big reinforcement learning efforts kind of all fizzled out—like post-Dota, as the reinforcement learning wasn't working out for solving those hard problems outside of the simulated game universes—then I started thinking, “Okay, maybe we kinda overshot there…”

But then GPT came out of nowhere and basically erased all that. It was like, "Oh, shit. Here's how RL is going to work. It's going to be the cherry on the cake. We're just going to focus on the cake for a while.” Now we have actually figured out a good recipe for baking a cake, which was not true before.

Before, it seemed like you were going to have to brute-force it end-to-end from the rewards. But now you can do the LeCun thing, of learning fast on generative models and then just doing a little bit of RL on top to make it do something specific.

Dwarkesh Patel
Now that you know that AGI is a thing that's coming, what’s your thinking around how you see your role in this timeline? How are you thinking about how to spend these next few years?

Gwern Branwen
I have been thinking about that quite a lot. What do I want to do? What would be useful to do?

I'm doing things now because I want to do them, regardless of whether it will be possible for an AI to do them in like 3 years. I do something because I want to. Because I like it, I find it funny or whatever. Or I think carefully about doing just the human part of it, like laying out a proposal for something.

If you take seriously the idea of getting AGI in a few years, you don't necessarily have to implement stuff and do it yourself. You can sketch out clearly what you want, and why it would be good and how to do it. And then just wait for the better AGI to come along and actually do it then. Unless there's some really compelling reason to do it right now and pay that cost of your scarce time.

But otherwise, I'm trying to write more about what is not recorded. Things like preferences and desires and evaluations and judgments. Things that an AI could not replace even in principle.

The way I like to put it is that "the AI cannot eat ice cream for you". It cannot decide for you which kind of ice cream you like. Only you can do that. And if anything else did, it would be worthless, because it's not your particular preference.

That's kind of the rubric. Is this something I want to do regardless of any future AI, because I enjoy it? Or is this something where I'm doing only the human part of it and the AGI can later on do it? Or is this writing down something that is unwritten and thus helping the future AI versions of me?

So if it doesn't fall under those 3, I have been trying to not do it.

If you look at it that way, many of the projects that people do now have basically no lasting value. They're doing things that they don't enjoy, which record nothing ephemeral of value that could not be inferred or generated later on. They are, at best, getting 2 or 3 years of utility out of it before it could have been done by an AI system.

Dwarkesh Patel
Wait, your timeline for when an AI could write a Gwern-quality essay is two to three years?

Gwern Branwen
Ehmm… I have ideas about how to make it possible, which might not require AGI if it combined my entire corpus. Many potential essay ideas are already mostly done in my corpus. So you don't need to be super intelligent to pull it out.

So let's talk about AGI in general: the Anthropic timeline of 2028 seems like a good personal planning starting point. Even if you're wrong, you probably weren't going to do a lot of projects within the next 3 years anyway. It's not like you really lost much by instead just writing down the description. You can always go back and do it yourself if you're wrong.

Dwarkesh Patel
You wrote an interesting comment about getting your work into the LLM training corpus: "there has never been a more vital hinge-y time to write."

Do you mean that in the sense that you will be this drop in the bucket that’s steering the Shoggoth one way or the other? Or do you mean it in the sense of making sure your values and persona persist somewhere in latent space?

Gwern Branwen
I mean both. By writing, you are voting on the future of the Shoggoth using one of the few currencies it acknowledges: tokens it has to predict. If you aren't writing, you are abdicating the future or your role in it. If you think it's enough to just be a good citizen, to vote for your favorite politician, to pick up litter and recycle, the future doesn't care about you.

There are ways to influence the Shoggoth more, but not many. If you don't already occupy a handful of key roles or work at a frontier lab, your influence rounds off to 0, far more than ever before. If there are values you have which are not expressed yet in text, if there are things you like or want, if they aren't reflected online, then to the AI they don't exist. That is dangerously close to won't exist.

But yes, you are also creating a sort of immortality for yourself personally. You aren't just creating a persona, you are creating your future self too. What self are you showing the LLMs, and how will they treat you in the future?

I give the example of Kevin Roose discovering that current LLMs—all of them, not just GPT-4—now mistreat him because of his interactions with Sydney, which "revealed" him to be a privacy-invading liar, and they know this whenever they interact with him or discuss him. Usually, when you use a LLM chatbot, it doesn't dislike you personally! On the flip side, it also means that you can try to write for the persona you would like to become, to mold yourself in the eyes of AI, and thereby help bootstrap yourself.

Dwarkesh Patel
Things like the Vesuvius Challenge show us that we can learn more about the past than we thought possible. They’ve leaked more bits of information that we can recover with new techniques. Apply that to the present and think about what the future superhuman intelligences will be trying to uncover about the current present. What kinds of information do you think are going to be totally inaccessible to the transhumanist historians of the future?

Gwern Branwen
Any kind of stable, long-term characteristics, the sort of thing you would still have even if you were hit on the head and had amnesia… Anything like that will be definitely recoverable from all the traces of your writing, assuming you're not pathologically private and destroy everything possible. That should all be recoverable.

What won't be recoverable will be everything that you could forget ordinarily: autobiographical information, how you felt at a particular time, what you thought of some movie. All of that is the sort of thing that vanishes and can't be recovered from traces afterwards.

If it wasn't written down, it wasn't written down.

Dwarkesh Patel
What is the biggest unresolved tension in your worldview?

Gwern Branwen
The thing I swing back and forth the most on is the relationship between human intelligence and neural network intelligence.

It's not clear in what sense they are two sides of the same coin, or one is an inferior version of the other. This is something that I constantly go back and forth on: “Humans are awesome.” “No, neural networks are awesome.” Or, “No, both suck.” Or, “Both are awesome, just in different ways.”

So every day I argue with myself a little bit about why each one is good or bad or how. What is the whole deal there with things like GPT-4 and memorization, but not being creative? Why do humans not remember anything, but we still seem to be so smart? One day I'll argue that language models are sample efficient compared to humans. The next day I'll be arguing the opposite.

Dwarkesh Patel
One of the interesting points you made to me last year was that AI might be the most polymathic topic to think about because there’s no field or discipline that is not relevant to thinking about AI. Obviously you need computer science and hardware. But you also need things like primatology and understanding what changed between chimp and human brains, or the ultimate laws of physics that will constrain future AI civilizations. That’s all relevant to understanding AI. I wonder if it’s because of this polymathic nature of thinking about AI. that you’ve been especially productive at it.

Gwern Branwen
I'm not sure it was necessary. When I think about others who were correct, like Shane Legg or Dario Amodei, they don't seem to be all that polymathic. They just have broad intellectual curiosity, broad general understanding, absolutely. But they’re not absurdly polymathic. Clearly you could get to the correct view without being polymathic. That's just how I happen to come to it at this point and the connection I’m making post hoc.

It wasn’t like I was using primatology to justify scaling to myself. It's more like I'm now using scaling to think about primatology. Because, obviously, if scaling is true, it has to tell us something about humans and monkeys and all other forms of intelligence. It just has to. If that works, it can't be a coincidence and totally unrelated. I refuse to believe that there are two totally unrelated kinds of intelligence, or paths to intelligence—where humans, monkeys, guppies, dogs are all one thing, and then neural networks and computers are another thing—and they have absolutely nothing to do with each other.

That's obviously wrong. They can be two sides of the same coin. They can obviously have obscure connections. Maybe one could be a better form or whatever. They can't just be completely unrelated. As if humans finally got to Mars and then simultaneously a bunch of space aliens landed on Mars for the first time and that's how we met. You would never believe that. It would be just too absurd.

Dwarkesh Patel
What is it that you are trying to maximize in your life?

Gwern Branwen
I maximize rabbit holes. I love more than anything else, falling into a new rabbit hole. That's what I really look forward to. Like this sudden new idea or area that I had no idea about, where I can suddenly fall into a rabbit hole for a while. Even things that might seem bad are a great excuse for falling into a rabbit hole.

Here’s one example. I buy some catnip for my cat and I waste $10 when I find out that he's catnip-immune. I can now fall into a rabbit hole of the question of “well, why are some cats catnip-immune? Is this a common thing in other countries? How does it differ in other countries? What alternative catnip drugs are there?” (It turned out to be quite a few.)

I was wondering, “How can I possibly predict which drug my cat would respond to? Why are they reacting in these different ways?”... Just a wonderful rabbit hole of new questions and topics I can master and get answers to, or create new ones, and exhaust my interest until I find the next rabbit hole I can dig and dive into.

Dwarkesh Patel
What is the longest rabbit hole you've gone on which didn't lead anywhere satisfying?

Gwern Branwen
That was my very old work on the anime Neon Genesis Evangelion, which I was very fond of when I was younger. I put a ludicrous amount of work into reading everything ever written about Evangelion in English and trying to understand its development and why it is the way it is. I never really got a solid answer on that before I burned out on it.

I actually do understand it now by sheer chance many years later. But at this point, I no longer care enough to write about it or try to redo it or finish it. In the end, it all wound up being basically a complete waste.

I have not used it in any of my other essays much at all. That was really one deep rabbit hole that I almost got to the end of, but I couldn't clinch it.

Dwarkesh Patel
How do you determine when to quit a rabbit hole? And how many rabbit holes do you concurrently have going on at the same time?

Gwern Branwen
You can only really explore two or three rabbit holes simultaneously. Otherwise, you aren't putting real effort into each one. You’re not really digging the hole, it's not really a rabbit hole. It's just something you are somewhat interested in. A rabbit hole is really obsessive. If you aren't obsessed with it and continually driven by it, it's not a rabbit hole. That’s my view. I’d say two or three max, if you're spending a lot of time and effort on each one and neglecting everything else.

As for when you exit a rabbit hole, you usually hit a very natural terminus where getting any further answers requires data that do not exist or you have questions that people don't know the answer to. You reach a point where everything dies out and you see no obvious next step.

One example would be when I was interested in analogs to nicotine that might be better than nicotine. That was a bit of a rabbit hole, but I quickly hit the dead end that there are none. That was a pretty definitive dead end. I couldn't get my hands on the metabolites of nicotine as an alternative. So if there are no analogs and you can't get your hands on the one interesting chemical you find, well that's that. That's a pretty definitive end to that rabbit hole.

Dwarkesh Patel
Have you always been the kind of person who falls into rabbit holes? When did this start?

Gwern Branwen
Oh, yeah. My parents could tell you all about that. I was very much your stereotypical nerdy little kid having the dinosaur phase and the construction equipment phase and the submarine and tank phase.

Dwarkesh Patel
Many kids are into “those things”, but they don't rabbit hole to the extent that they’re forming taxonomies about the different submarines and flora and fauna and dinosaurs, and developing theories of why they came to be and so forth.

Gwern Branwen
Well, I think it's more that people grow out of being very into rabbit holes as a kid. For me, it was not so much that I was all that exceptional in having obsessions as a kid.

It’s more that they never really stopped. The tank phase would be replaced by my Alcatraz phase where I would go to the public library and check out everything they had about Alcatraz. That would be replaced by another phase where I was obsessed with ancient Japanese literature. I would check out everything that the library had about Japanese literature before the haiku era. The process of falling into these obsessions kept going for me.

Dwarkesh Patel
By the way, do you mind if I ask how long you’ve been hearing impaired?

Gwern Branwen
Since birth. I've always been hearing impaired.

Dwarkesh Patel
And I assume that impacted you through your childhood and at school?

Gwern Branwen
Oh, yeah, absolutely, hugely. I went to a special ed school before kindergarten for hearing impaired and other handicapped kids. During school it was very rough because at the time, we had to use pairs of hearing aids hooked up to the teacher. Every class I would have to go up to the teacher with a big brown box with the hearing aids so she could use it. I always felt very humiliated by that, how it marked me out as different from other kids, not being able to hear.

The effects on socializing with other kids is terrible because you're always a second behind in conversation if you're trying to understand what the other person is saying. The hearing aids back then were pretty terrible. They've gotten a lot better but back then they were pretty terrible. You would always be behind. You'd always be feeling like the odd person out. Even if you could have been a wonderful conversationalist, you can't be if you're always a second behind and jumping in late. When you are hearing impaired, you understand acutely how quickly conversation moves. Milliseconds separate the moment between jumping in and everyone letting you talk, and someone else talking over you. That's just an awful experience if you're a kid who's already kind of introverted. It’s not like I was very extroverted as a kid, or now. So that was always a barrier.

Then you had a lot of minor distortions. I still have a weird fear of rain and water because it was drilled into me that I could not get the hearing aids wet because they were very expensive. I would always feel a kind of low-grade, stressful anxiety around anywhere like a pool, a body of water. Even now, I always feel weird about swimming, which I kind of enjoy. But I'm always thinking to myself, “Oh, wow, I won't be able to see because I'm nearsighted and I won't be able to hear because I had to take off my hearing aid to go in. I can't hear anything that anyone says to me in the pool, which takes a lot of the fun out of it.”

Dwarkesh Patel
You have a list of open questions on your website and one of them is, “Why do the biographies of so many great people start off with traumatic childhoods?” I wonder if you have an answer for yourself. Was there something about the effect that hearing impairment had on your childhood, your inability to socialize, that was somehow important to you becoming Gwern?

Gwern Branwen
It definitely led to me being so much of a bookworm. That's one of the things you can do as a kid which is completely unaffected by any kind of hearing impairment. It was also just a way to get words and language. Even now, I still often speak words in an incorrect way because I only learned them from books. It's the classic thing where you mispronounce a word because you learn it from a book and not from hearing other people sound it out and say it.

Dwarkesh Patel
Is your speech connected to your hearing impairment?

Gwern Branwen
Yes. The deaf accent is from the hearing impairment. It's funny, at least three people on this trip to SF have already asked me where I am really from. It's very funny. You look at me and you’re like, “Oh, yes, he looks like a perfectly ordinary American.” Then I open my mouth and it’s, “Oh, gosh, he's Swedish. Wow. Or maybe possibly Norwegian. I'll ask him where he's actually from. How did he come to America?”

I've been here the whole time! That's just how hearing impaired people sound. No matter how fluent you get, you still bear the scars of growing up hearing impaired. At least when you're born with it—or from very early childhood—your cognitive development of hearing and speech is always a little off, even with therapy.

One reason I don't like doing podcasts is that I have no confidence that I sound good, or at least, sound nearly as good as I write. Maybe I'll put it that way.

Dwarkesh Patel
What were you doing with all these rabbit holes before you started blogging? Was there a place where you would compile them?

Gwern Branwen
Before I started blogging, I was editing Wikipedia.

That was really gwern.net before gwern.net. Everything I do now with my site, I would have done on English Wikipedia. If you go and read some of the articles I am still very proud of—like the Wikipedia article on Fujiwara no Teika—and you would think pretty quickly to yourself, “Ah yes, Gwern wrote this, didn't he?”

Dwarkesh Patel
Is it fair to say that the training that required to make gwern.net happened on Wikipedia?

Gwern Branwen
Yeah. I think so. I have learned far more from editing Wikipedia than I learned from any of my school or college training. Everything I learned about writing I learned by editing Wikipedia.

Dwarkesh Patel
Honestly, it sounds like Wikipedia is a great training ground if you wanted to make a thousand more Gwerns. This is where we train them.

Gwern Branwen
Building something like an alternative to Wikipedia could be a good training ground. For me it was beneficial to combine rabbit-holing with Wikipedia, because Wikipedia would generally not have many good articles on the thing that I was rabbit-holing on.

It was a very natural progression from the relatively passive experience of rabbit-holing—where you just read everything you can about a topic—to compiling that and synthesizing it on Wikipedia. You go from piecemeal, a little bit here and there, to writing full articles. Once you are able to write good full Wikipedia articles and summarize all your work, now you can go off on your own and pursue entirely different kinds of writing now that you have learned to complete things and get them across the finish line.

It would be difficult to do that with the current English Wikipedia. It's objectively just a much larger Wikipedia than it was back in like 2004. But not only are there far more articles filled in at this point, the editing community is also much more hostile to content contribution, particularly very detailed, obsessive, rabbit hole-y kind of research projects. They would just delete it or tell you that this is not for original research or that you're not using approved sources. Possibly you’d have someone who just decided to get their jollies that day by deleting large swathes of your specific articles. That of course is going to make you very angry and make you probably want to quit and leave before you get going.

So I don't quite know how you would figure out this alternative to Wikipedia, one that empowers the rabbit holer as much as the old Wikipedia did.

When you are an editor with Wikipedia, you have a very empowered attitude because you know that anything in it could be wrong and you could be the one to fix it. If you see something that doesn't make sense to you, that could be an opportunity for an edit.

That was, at least, the Wiki attitude: anyone could fix it, and “anyone” includes you.

Dwarkesh Patel
When you were an editor on Wikipedia, was that your full-time occupation?

Gwern Branwen
It would eat as much time as I let it. I could easily spend 8 hours a day reviewing edits and improving articles while I was rabbit-holing. But otherwise I would just neglect it and only review the most suspicious diffs on articles that I was particularly interested in on my watchlist. I might only spend like 20 minutes a day. It was sort of like going through morning email.

Dwarkesh Patel
Was this while you were at university or after?

Gwern Branwen
I got started on Wikipedia in late middle school or possibly early high school.

It was kind of funny. I started skipping lunch in the cafeteria and just going to the computer lab in the library and alternating between Neopets and Wikipedia. I had Neopets in one tab and my Wikipedia watch lists in the other.

Dwarkesh Patel
Were there other kids in middle school or high school who were into this kind of stuff?

Gwern Branwen
No, I think I was the only editor there, except for the occasional jerks who would vandalize Wikipedia. I would know that because I would check the IP to see what edits were coming from the school library IP addresses. Kids being kids thought they would be jerks and vandalize Wikipedia.

For a while it was kind of trendy. Early on, Wikipedia was breaking through to mass awareness and controversy. It's like the way LLMs are now. A teacher might say, "My student keeps reading Wikipedia and relying on it. How can it be trusted?"

So in that period, it was kind of trendy to vandalize Wikipedia and show your friends. There were other Wikipedia editors at my school in that sense, but as far as I knew I was the only one building it, rather than wrecking it.

Dwarkesh Patel
When did you start blogging on gwern.net? I assume this was after the Wikipedia editor phase. Was that after university?

Gwern Branwen
It was afterwards. I had graduated and the Wikipedia community had been very slowly moving in a direction I did not like. It was triggered by the Siegenthaler incident which I feel was really the defining moment in the trend toward deletionism on Wikipedia. It just became ever more obvious that Wikipedia was not the site I had joined and loved to edit and rabbit hole on and fill in, and that if I continued contributing I was often just wasting my effort.

I began thinking about writing more on my own account and moving into non-Wikipedia sorts of writings: persuasive essays, nonfiction, commenting, or possibly even fiction. I began gently moving beyond things like Reddit and LessWrong comments to start something longform.

Dwarkesh Patel
What was your first big hit?

Gwern Branwen
Silk Road. I had been a little bit interested in Bitcoin, but not too seriously interested in it because it was not obvious to me that it was going to work out, or even was technologically feasible. But when Adrian Chen wrote his Gawker article about buying LSD off Silk Road, all of a sudden I did a complete 180. I had this moment of, "Holy shit, this is so real that you can buy drugs off the Internet with it!"

I looked into the Chen article and it was very obvious to me that people wanted to know what the ordering process was like. They wanted more details about what it's like, because the article was very brief about that. It didn't go into any real detail about the process.

So I thought, "Okay, I'm interested in nootropics. I'm interested in drugs. I will go and use Silk Road. I will document it for everyone, instead of everyone pussyfooting around it online and saying, 'Oh, a friend of mine ordered off Silk Road and it worked.' None of that bullshit. I will just document it straightforwardly."

I ordered some Adderall, I think it was, and documented the entire process with screenshots. I wrote it up and wrote some more on the intellectual background. That was a huge hit when I published it. It was hundreds of thousands of hits. It's crazy. Even today when I go to the Google Analytics charts, you can still see "Silk Road" spiking vertically like crazy and then falling back down. Nothing else really comes near it in terms of traffic. That was really quite something, to see things go viral like that.

Dwarkesh Patel
What are the counterfactual career trajectories and life paths that could have been for you if you didn't become an online writer? What might you be doing instead that seems plausible?

Gwern Branwen
I could definitely have been an AI researcher, or possibly in management at one of the big AI companies. I would have regretted not being able to write about stuff, but I would've taken satisfaction in making it happen and putting my thumbprint on it. Those are totally plausible counterfactuals.

Dwarkesh Patel
Why didn't you?

Gwern Branwen
I kind of fell off that track very early on in my career when I found the curriculum of Java to be excruciatingly boring and painful. So I dropped out of computer science. That kind of put me off that track early on.

And then various early writing topics made it hard to transition in any other way than starting a startup, which I'm not really temperamentally suited for. Things like writing about the darknet markets or behavioral genetics, these are topics which don't exactly scream "great hire."

Dwarkesh Patel
Has agency turned out to be harder than you might have thought initially? We have models that seem like they should be able to do all of the individual things that a software engineer does. For example, all the code they might write, all the individual pull requests. But it seems like a really hard problem to get them to act as a coherent, autonomous, software engineer that puts in his eight hours a day.

Gwern Branwen
I think agency is, in many senses, actually easier to learn than we would have thought ten years ago. But we actually aren't learning agency at all in current systems. There’s no selection for that. All the agency there is is an accidental byproduct of somebody training on data.

So from that perspective, it's miraculous that you can ask an LLM to try to do all these things and they have a non-trivial success rate. If you told people ten years ago—that you could just behavior-clone on individual letters following one by one, and you could get coherent action out of it and control robots and write entire programs—their jaws would drop and they would say that you've been huffing too many fumes from DeepMind or something.

The reason that agency doesn't work is that we do so little actual agency training at all. An example of how you would do agency directly would be like Gato from DeepMind. There they’re actually training agents. Instead we train them on Internet scrapes which merely encode the outputs of agents or occasional descriptions of agents doing things. There’s no actual logging of state/action/result/reward sequences like a proper reinforcement learning setup would have.

I would say that what's more interesting is that nobody wants to train agents in a proper reinforcement learning way. Instead, everyone wants to train LLMs and do everything with as little RL as possible in the backend.

Dwarkesh Patel
What would a person like you be doing before the Internet existed?

Gwern Branwen
If the Internet did not exist, I would have to have tried to make it in regular academia and maybe narrow my interests a lot more, something I could publish on regularly.

Or I could possibly have tried to opt out and become a librarian like one of my favorite writers, Jorge Luis Borges. He was a librarian until he succeeded as a writer. Of course, I've always agreed with him about imagining paradise as a kind of library. I love libraries.

I regret that all the reading I do is now on the computer and I don't get to spend much time in physical libraries. I do genuinely love them, just pouring through the stacks and looking for random stuff. Some of the best times for me in university was being able to go through these gigantic stacks of all sorts of obscure books and just looking at a random spine, pulling stuff off the shelf and reading obscure, old technical journals to see all the strange and wonderful things they were doing back then, which now have been forgotten.

Dwarkesh Patel
If you could ask Borges one question, what would it be?

Gwern Branwen
Oh. He's a real hero of mine. This is not something I want to give a bad answer to.

[“Would it have been worth living if you could never write, only read, like the people in ‘The Library of Babel’?”]

Dwarkesh Patel
Can I ask why he's a hero of yours?

Gwern Branwen
When I was younger, one of the science fiction books that really impressed me was Dan Simmons' Hyperion, especially The Fall of Hyperion. In there, he alludes to Kevin Kelly's Out of Control book, which strongly features the parable of “The Library of Babel.” From there, I got the collected editions of Borges’ fiction and nonfiction. I just read through them again and again.

I was blown away by the fact that you could be so creative, with all this polymathic knowledge and erudition, and write these wonderful, entertaining, provocative short stories and essays. I thought to myself, “If I could be like any writer, any writer at all, I would not mind being Borges.”

Dwarkesh Patel
Borges has a short poem called "Borges and I" where he talks about how he doesn’t identify with the version of himself that is actually doing the writing and publishing all of this great work. I don’t know if you identify with that at all.

Gwern Branwen
When I was a kid, I did not understand that essay, but I think I understand it now.

Dwarkesh Patel
What are other pieces of either literature that you encountered where now you really understand what they were getting at but you didn’t when you first came across them?

Gwern Branwen
Ted Chiang's "Story of Your Life". I completely blew [it] understanding it the first time I read it. I had to get a lot more context where I could actually go back and understand what his point was. Gene Wolfe's "Suzanne Delage" story was a complete mystery to me. It took like 14 years to actually understand it. But I'm very proud of that one.

Dwarkesh Patel
What did you figure out about Suzanne Delage?

Gwern Branwen
Gene Wolfe's "Suzanne Delage" is a very, very short story about a guy remembering not meeting a woman in his local town and thinking, “Oh, that's kind of strange.” That's the whole story. Nobody has any idea what it means, even though we're told that it means something. Gene Wolfe is a genius writer, but nobody could figure it out for like 40 years.

Last year I figured it out. It turns out it's actually a subtle retelling of Dracula, where Dracula invades the town and steals the woman from him. He's been brainwashed by Dracula—in a very Bram Stoker way—to forget it all. Every single part of the story is told by what's not said in the narrator's recollection. It's incredible. It's the only story I know which is so convincingly written by what's not in it.

Dwarkesh Patel
That’s crazy that you figured that out. The Ted Chiang story, the “Story of Your Life”, can you remind me what that one’s about?

Gwern Branwen
The surface story is just about a bunch of weird aliens who came to Earth.

Dwarkesh Patel
Oh, that's right, yeah. It’s the same plot as Arrival.

Gwern Branwen
They had a weird language which didn't have a sense of time. The narrator learned to see the future, and then the aliens left.

Dwarkesh Patel
What is it that you realized about that story?

Gwern Branwen
The first time I read it, it struck me as just a kind of stupid ESP story about seeing the future, very stupid, boring, standard conventional, verbose, and dragging in much irrelevant physics. Only a while after that did I understand that it was not about time travel or being able to see the future.

It was instead about a totally alien kind of mind that’s equally valid in its own way, in which you see everything as part of an already determined story heading to a predestined end. This turned out to be mathematically equivalent and equally powerful as our conventional view of the world—events marching one by one to an unknown and changing future.

That was a case where Chiang was just writing at too high a level for me to understand. I pattern-matched it to some much more common, stupid story.

Dwarkesh Patel
How do you think about the value of reading fiction versus nonfiction?

Gwern Branwen
You could definitely spend the rest of your life reading fiction and not benefit whatsoever from it other than having memorized a lot of trivia about things that people made up.

I tend to be pretty cynical about the benefits of fiction. Most fiction is not written to make you better in any way. It's written just to entertain you, or to exist and to fill up time.

Dwarkesh Patel
But it sounds like your own ideas have benefited a lot from the sci-fi that you read.

Gwern Branwen
Yeah, but it's extremely little sci-fi. Easily 99% of the sci-fi I read was completely useless to me. I could have easily cut it down to 20 novels or short stories which actually were good enough and insightful enough to actually change my view. One volume of Blindsight by Peter Watts is worth all hundred Xanth novels, or all 500 Expanded Universe novels of Star Wars.

Dwarkesh Patel
The ones that you did find insightful, the top 20 or so, what did they have in common?

Gwern Branwen
I would say that the characteristic they have is taking non-human intelligence seriously.

It doesn't have to be artificial intelligence necessarily. It's taking the idea of non-human intelligence seriously and not imagining your classic sci-fi scenario of humans going out into the galaxy with rayguns—the sort of thing where you have rockets and rayguns but you don't have cell phones.

People complain that the Singularity is a sort of boring, overused sci-fi trope. But if you went out and actually grabbed random books of science fiction, you would find that less than 1% contain anything remotely like that, or have any kind of relevance to the current context that we actually face with AI.

Dwarkesh Patel
Do people tend to underestimate or overestimate your intelligence?

Gwern Branwen
I would say they overestimate it. They mistake for intelligence the fact that I remember many things, that I have written many things over many years. They imagine that if they sat me down, I could do it all spontaneously at the moment that they're talking to me. But with many things I have thought about, I have the advantage of having looked at things before. So I'm cheating. When I talk to people, I may just be quoting something I've already written, or at least thought about.

So I come off as a lot smarter than I actually am. I would say I'm not really all that smart, compared to many people I've known, who update very fast on the fly. But in the end, it's the output that matters, right?

Dwarkesh Patel
I guess there is an on-the-fly intelligence. But there's another kind too which is this ability to synthesize things over a long period of time, and then come up with grand theories as a result of these different things that you're seeing. I don't think that's just crystallized intelligence, right?

Gwern Branwen
It's not just crystallized intelligence, but if you could see all the individual steps in my process, you'd be a lot less impressed. If you could see all of the times I just note down something like, "Hmm, that's funny." Or, "Huh, another example of that," and if you just saw each particular step, you would say that what I was doing was reasonable and not some huge sign of brilliance. It would make sense to you in that moment. It's only when that happens over a decade, and you don't see the individual stuff, that my output at the end looks like magic.

One of my favorite quotes about this process is from the magicians Penn & Teller. Teller says "magic is putting in more effort than any reasonable person would expect you to." He tells a story about how they make cockroaches appear from a top hat. The trick is that they researched and found special cockroaches, and then found special styrofoam to trap the cockroaches, and arranged all that, for just a single trick. No reasonable person would do that, but they did because they wanted the trick to really pay off. The result is cockroaches somehow appearing from an empty hat.

If you could see each step, it would make sense on its own, it would just look effortful. But when you see only the final trick, then that whole process and its output becomes magic.

Dwarkesh Patel
That's one of the interesting things about your process. There are a couple of writers like Matt Levine or Byrne Hobart who write an article every day. I think of them almost like autoregressive models. For you, on some of the blog posts you can see the start date and

end date that you list on your website of when you've been working on a piece. Sometimes it's like 2009 to 2024. I feel like that's much more like diffusion. You just keep iterating on the same image again and again.

One of my favorite blog posts of yours is "Evolution as Backstop for RL," where you talk about evolution as basically a mechanism to learn a better learning process. And that explains why corporations don't improve over time but biological organisms do. I'm curious if you can walk me through the years that it took to write that. What was that process like, step by step?

Gwern Branwen
So the "Backstop" essay that you're referring to is the synthesis of seeing the same pattern show up again and again: a stupid, inefficient way of learning, which you use to learn something smarter, but where you still can't get rid of the original one entirely.

Sometimes examples would just connect to each other when I was thinking about this. Other times —when I started watching for this pattern—I would say, "Oh yes, 'pain' is a good example of this. Maybe this explains why we have pain in the very specific way that we have, when you can logically imagine other kinds of pain, and those other pains would be smarter, but nothing keeps them honest."

So you just chain them one by one, these individual examples of the pattern, and just keep clarifying the central idea as you go. Wittgenstein says that you can look at an idea from many directions and then go in spirals around it. In an essay like "Backstop," it's me spiraling around this idea of having many layers of "learning" all the way down.

Dwarkesh Patel
Once you notice one example of this pattern, like this pain example, do you just keep adding examples to that? Walk me through the process over time.

Gwern Branwen
For that specific essay, the first versions were about corporations not evolving. Then, as I read more and more of the meta reinforcement learning literature, from DeepMind especially, I added in material about neural networks. I kept reading and thinking about the philosophy of mind papers that I had read. I eventually nailed down the idea that pain might be another instance of this: "Pain makes us learn. We can't get rid of it, because we need it to keep us honest." At that point you have more or less the structure of the current essay.

Dwarkesh Patel
Are there examples where it's not a matter of accumulating different instances of what you later realize is one bigger pattern? Rather, you just have to have the full thesis at once.

Gwern Branwen
For those essays where there is an individual eureka moment, there's usually a bunch of disparate things that I have been making notes on that I don't even realize are connected. They just bother me for a long time. They sit there bothering me. I keep looking for explanations for each one and not finding them. It keeps bothering me and bothering me.

One day, I hit something that suddenly makes me go, “Bam, eureka. These are all connected!” Then I just have to sit down and write a single gigantic essay that pours out about it and then it's done. That particular essay will be done at that point—right in one go. I might add in many links to it or references later on, but it will not fundamentally change.

Dwarkesh Patel
What's an example of an essay that had this process?

Gwern Branwen
Someone asked about how I came up with one yesterday, as a matter of fact. It’s one of my oldest essays, “The Melancholy of Subculture Society.”

For that one, I had been reading miscellaneous things like David Foster Wallace on tennis, people on Internet media like video games. One day it just hit me: it's incredibly sad that we have all these subcultures and tribes online that can find community together, but they are still incredibly isolated from the larger society. One day, a flash just hit me about how beautiful and yet also sad this is.

I sat down and wrote down the entire thing more or less. I've not really changed it all that much. I've added more links and quotes and examples over time, but nothing important. The essence was just a flash and I wrote it down while it was there.

Dwarkesh Patel
One of the interesting quotes you have in the essay is from David Foster Wallace when he’s talking about the tennis player Michael Joyce. He’s talking about the sacrifices Michael Joyce has had to make in order to be top ten in the world at tennis. He’s functionally illiterate because he’s been playing tennis every single day since he was seven or something, and not really having any life outside of tennis.

What are the Michael Joyce-type sacrifices that you have had to make to be Gwern?

Gwern Branwen
That's a hard hitting question, Dwarkesh! “How have I amputated my life in order to write?”… I think I've amputated my life in many respects professionally and personally, especially in terms of travel. There are many people I envy for their ability to travel and socialize, or for their power and their positions in places like Anthropic where they are the insiders. I have

sacrificed whatever career I could have had, or whatever fun lifestyle: a digital nomad lifestyle and going outdoors, being a Buddhist monk, or maybe a fancy trader. All those have had to be sacrificed for the patient work of sitting down every day and reading papers until my eyes bleed, and hoping that something good comes out of it someday.

Dwarkesh Patel
Why does it feel like there's a trade off between the two? There are obviously many writers who travel a lot like Tyler Cowen. There are writers who have a lot of influence such as Jack Clark at Anthropic. Why does it feel like you can't do both at the same time?

Gwern Branwen
I can't be or be compared to Tyler Cowen. Tyler Cowen is a one-man industry.

Dwarkesh Patel
So is Gwern.

Gwern Branwen
Yeah, but he cannot be replicated. I just cannot be Tyler Cowen. Jack Clark, he is also his own thing. He's able to write the stories in his issues very well while also being a policy person. I respect them and admire them.

But none of those quite hit my particular interest and niche at following weird topics for a long period of time, and then collating and sorting through information. That requires a large commitment to reading vast masses of things in the hopes that some tiny detail perhaps will turn out to one day be important.

Dwarkesh Patel
So walk me through this process. You talked about reading papers until your eyes bleed at the end of the day. You wake up in the morning and you go straight to the papers? What does your day look like?

Gwern Branwen
The workflow right now is more like: I wake up, I do normal morning things, and then I clean up the previous day's work on the website. I deal with various issues, like formatting or spelling errors. I review it and think if I properly collated everything and put it in the right places. Sometimes I might have an extra thought that I need to add in or make a comment that I realize was important. That's the first step.

After that, I often will shamelessly go to Twitter or my RSS feed and just read a large amount until perhaps I get distracted by a comment or a question from someone and maybe do some writing on that.

After that, I take a break for lunch or whatever, and then go back to that and just keep going at it. Somewhere around evening, I will often get exhausted from all that, and try to do a real project or contribution to something. I'll actually sit down and work on whatever I'm supposed to be working on that day.

After that, I would typically go to the gym. By that point, I really am burned out from everything. Yes, I like going to the gym—not because I'm any kind of meathead or athlete or even really enjoy weightlifting—but because it's the most diametrically opposite thing I can do to sitting in front of a computer.

Dwarkesh Patel
This is your theory of burnout, right? That you have to do the exact opposite?

Gwern Branwen
Yes, when people experience burnout, you just feel a lack of reward for what you're doing or what you're working on. You just need to do something different. Something as different as possible. Maybe you could do better than weightlifting, but it does feel very different from anything I do in front of a computer.

Dwarkesh Patel
I want to go back to your process. Everyday, you're loading up all this context. You're reading all the RSS feeds and all these papers. Are you basically making contributions to all your essays, adding a little bit here and there every single day? Or are you building up some potential which will manifest itself later on as a full essay, a fully formed thesis?

Gwern Branwen
I would say it's the latter one. All the minor low-level additions and pruning and fixing I do is really not that important. It's more just a way to make nicer essays. It's a purely aesthetic goal, to make as nice an essay as I possibly can. I'm really waiting to see what happens next. What will be the next thing I'll be provoked to write about? It's just passing the time in between sudden eruptions.

I feel that for many writers, you can't neglect the gardening process. You don't harvest every day. You have to tend the garden for a long time in between harvests. If you start to neglect the gardening because you're gallivanting around the world... Let's say you're going to book signing events and doing all the publicity stuff. Then you're not doing the work of being in there and tending your garden. That's undermining your future harvest, even if you can't see it right now.

If you ask what is Tyler Cowen's secret to being Tyler Cowen, my guess would be that he's just really good at tending his garden, even as he travels a crazy amount. That would be his secret, that he's able to read books on a plane. I can't read books on a plane. He's able to write everything in the airport. I can do a little bit of writing in the airport but not very much. He's just very robust to the wear and tear of traveling. I'll be collapsing in the hotel room after talking to people for eight hours. He's able to talk to people for eight hours and then go do podcasts and talk to someone for another four hours! That's extremely admirable, but I just can't do that.

Dwarkesh Patel
How often do you get bored? It sounds like you're spending your whole day reading different things. Are they all just inherently interesting to you? Or do you just trudge through it even when it's not compelling to you in the moment?

Gwern Branwen
I don't think I get bored too easily because I switch between so many different topics. Even if I'm kind of sick of deep learning papers, well, I have tons of other things I can read or argue with people about. So I don't really get bored. I just get exhausted. I have to go off and do something else, like lift weights.

Dwarkesh Patel
What is your most unusual but successful work habit?

Gwern Branwen
I think I get a lot more mileage out of arguing with people online than... pretty much any other writer does. [Patel laughs] Hey, I'm trying to give a genuine answer here, not some stupid thing about note-taking—a real answer!

I get a lot more out of arguing with people than most people do. You need motivation to write and actually sit down, and crystallize something and do the harvest. After you tend your garden, you do have to do the harvest, and the harvest can be hard work. It's very tedious.

There are many people I talk to who have many great ideas. But they don't want to harvest because it's tedious and boring. And it's very hot out there in the fields, reaping. You're getting dusty and sweaty. Why wouldn't you just be inside having lemonade?

But motivation from arguing and being angry at people online is in plentiful supply. So I get a lot of mileage out of people being wrong on the Internet.

Dwarkesh Patel
What are the pitfalls of an isolated working process?

Gwern Branwen
There’s the obvious one: you could be arbitrarily wrong when writing by yourself and just become a crazy loony by having a ‘big take’.

Aside from that, you also have the issue of the emotional toll of not having colleagues that you can convince. You often just have the experience of shouting onto the internet that continues to be wrong despite your shouting.

One thing I observe is that very often independent writers are overcome by resentment and anger and disappointment. They sort of spiral out into bitterness and crankdom from there. That's kind of what kills them. They could have continued if they’d only been able to let go of the ideas and arguments and move on to the next topic.

So I say that ‘spite can be a great motivation to write, but you have to use it skillfully and let it go afterwards’. You can only have it while you need motivation to write. If you keep going and hold on to it, you're poisoning yourself.

Dwarkesh Patel
I'm sure you're aware that many people comment on the fact that ‘if Gwern put the effort he spends optimizing the CSS on his website towards more projects and more writing, the benefits to society could be measured in the nearest million dollars’. What's your reaction to people who say you're spending too much time on site design?

Gwern Branwen
I have no defense at all there in terms of objective benefits to society. I do it because I'm selfish and I like it. That is my defense. I like the aesthetics of my website and it is a hobby.

Dwarkesh Patel
Does the design help you think?

Gwern Branwen
It does because I like rereading my stuff more when I can appreciate the aesthetics of it and the beauty of the website. It’s easier for me to tolerate reading something for the hundredth time when I would otherwise be sick to death of it. Site maintenance for the author is inherently this kind of spaced repetition. If I go over pages to check that some new formatting feature worked, I am getting spaced repetition there. More than once, I’ve gone to check some stupid CSS issue and looked at something and thought, “Oh, I should change something,” or, “Oh, that means something.”

So in a way, it's not as much of a waste as it looks, but I can't defend it entirely. If someone wants to make their own website, they should not invest that much for the aesthetic value.

I just want a really nice website. There's so many bad websites out there that it depresses me. There's at least one website I love.

Dwarkesh Patel
By the way, I'm going to mention this since you never mentioned it yourself. But I think the main way you fund your research is through your Patreon, right? You never advertise it but I feel—with the kind of thing you're doing—if it were financially viable and got adequate funding, not only would you be able to keep doing it but other people who wanted to be independent researchers could see it's a thing you can do. It's a viable thing you can do. More Gwerns would exist.

Gwern Branwen
Well, I don't necessarily want more Gwerns to exist. I just want more writers and more activeness and more agency in general.

I would be perfectly happy if someone simply wrote more Reddit comments and never took a dollar for their writings and just wrote better Reddit comments. I'd be perfectly happy if someone had a blog and they kept writing, but they just put a little more thought into the design. I'd be perfectly happy if no one ever wrote something, but they hosted PDFs so that links didn't rot.

In general, you don't have to be a writer delivering longform essays. That's just one of many ways to write. It happened to be the one that I personally kind of prefer. But it'd be totally valid to be a Twitter thread writer.

Dwarkesh Patel
How do you sustain yourself while writing full time?

Gwern Branwen
Patreon and savings. I have a Patreon which does around $900-$1000/month, and then I cover the rest with my savings. I got lucky with having some early Bitcoins and made enough to write for a long time, but not forever. So I try to spend as little as possible to make it last.

I should probably advertise the Patreon more, but I'm too proud to shill it harder.
It's also awkward trying to come up with some good rewards which don't entail a paywall. Patreon and Substack work well for a lot of people like Scott Alexander, because they like writing regular newsletter-style updates but I don't like to. I just let it run and hope it works.

Dwarkesh Patel
Wait if you're doing $900-1000/month and you're sustaining yourself on that, that must mean you're sustaining yourself on less than $12,000 a year. What is your lifestyle like at $12K?

Gwern Branwen
I live in the middle of nowhere. I don't travel much, or eat out, or have health insurance, or anything like that. I cook my own food. I use a free gym. There was this time when the floor of my bedroom began collapsing. It was so old that the humidity had decayed the wood. We just got a bunch of scrap wood and a joist and propped it up. If it lets in some bugs, oh well! I live like a grad student, but with better ramen. I don't mind it much since I spend all my time reading anyway.

Dwarkesh Patel
It's still surprising to me that you can make rent, take care of your cat, deal with any emergencies, all of that on $12K a year.

Gwern Branwen
I'm lucky enough to be in excellent health and to have had no real emergencies to date. This can't last forever, and so it won't. I'm definitely not trying to claim that this is any kind of ideal lifestyle, or that anyone else could or should try to replicate my approach! I got lucky with Bitcoin and with being satisfied with living like a monk and with my health.

Anyone who would like to take up a career as a writer or blogger should understand that this is not an example they can imitate. I'm not trying to be a role model.

Every writer will have to figure it out a different way. Maybe it can be something like a Substack, or just writing on the side while slinging Javascript for a tech company. I don't know.

Dwarkesh Patel
It seems like you've enjoyed this recent trip to San Francisco? What would it take to get you to move here?

Gwern Branwen
Yeah, it is mostly just money stopping me at this point. I probably should bite the bullet and move anyway. But I'm a miser at heart and I hate thinking of how many months of writing runway I'd have to give up for each month in San Francisco.

If someone wanted to give me, I don't know, $50–100K/year to move to SF and continue writing full-time like I do now, I'd take it in a heartbeat. Until then, I'm still trying to psych myself up into a move.

Dwarkesh Patel
That sounds very doable. If somebody did want to contribute to making this move, and your research more generally, possible, how would they get in touch with you?

Gwern Branwen
I have a Stripe donation page, or they could just email me at gwern@gwern.net.

Dwarkesh Patel
By when will AI models be more diverse than the human population?

Gwern Branwen
I'm going to say that if you exclude capability from that, AI models are already much more diverse cognitively than humans are.

Different LLMs think in very distinct ways that you can tell right away from a sample of them. An LLM operates nothing like a GAN. A GAN also is totally different from VAEs. They have totally different latent spaces, especially in the lower end, where they’re small or bad models. They have wildly different artifacts and errors in a way that we would not see with humans.

Humans are really very quite similar in writing and attitude compared to these absurd outputs of different kinds of models.

Dwarkesh Patel
Really? If you look at Chatbot Arena and you see side-by-side comparisons of the outputs of different models, it's often very hard to tell which ones comes from which model.

Gwern Branwen
Yeah but this is all very heavily tuned. Now you're restricting it to relatively recent LLMs, with everyone riding each other's coattails and often training on the same exact data. This is a situation much closer to if they were identical twins.

If I don't restrict myself to just LLMs and I compare the wide diversity of say image generation models, they often have totally different ways. Some of them seem as similar to each other as ants do to beavers.

Within LLMs, I would agree that there has been a massive loss of diversity. Things used to be way more diverse among LLMs. But across deep learning in general, we’ve seen a whole range of minds and ways to think that you won't find in any philosophy of mind paper.

Dwarkesh Patel
What's an example of two models that have these sorts of cognitive differences?

Gwern Branwen
I’ll give one example I was telling someone the other day. GAN models have incentives to hide things because it's an adversarial loss, whereas diffusion models have no such thing. So GAN models are ‘scared’. They put ‘hands’ off the screen. They just can't think about hands. Whereas diffusion models think about hands, but in their gigantic, monstrous, Cthulhu-esque abortions.

Dwarkesh Patel
People weren't paying attention to scaling in 2020. Is there some trend today where people aren’t really comprehending the full implications of where this is headed?

Gwern Branwen
I'm excited by the weight-loss drugs, the GLP drugs. Their effects in general on health and addiction across all sorts of behaviors really surprised me. No one predicted that as far as I know. While the results are still very preliminary, it does seem like it's real.

I think that’s going to tell us something important about human willpower and dysfunctionality. What's going wrong broadly in the modern environment?

Dwarkesh Patel
Do GLP drugs break the Algernon argument—the one you listed in your blog post—that if there are any simple and useful interventions without bad side effects, then evolution should have already found them?

Gwern Branwen
It's too soon to say because we haven't actually figured out what's going on with the GLPs to even understand what they are doing at all, what has the off target. It's kind of crazy that activating and deactivating both work.

It's a completely crazy situation. I don't really know what to think about the Algernon argument there. It could be that the benefits actually decrease fitness in the fertility sense because you're going out and having a happy life instead of having kids. No offense to parents. Or it could just be that it's hitting the body in a way that's really, really hard to replicate in any kind of genetic way. Or it could be that it's just too soon.

When I think back, I see that the obesity crisis only became obvious around the 1990s. It's quite recent. I look back at photos and today is completely unrecognizable from 1990. You look at photos and people are still thin. You look at photos now and everyone is like a blimp. So you can't possibly have any kind of Algernon argument over 20 or 30 years.

Dwarkesh Patel
When you look back at the Romans and you see how lead was constantly poisoning the entire city, what credence do you give to the possibility that something in our environment is having an effect on us on a similar magnitude of what lead was doing to the ancient Romans?

Gwern Branwen
I think the odds of there being something as bad as lead is almost 100%. We have so many things out there. Chemists are always cooking up new stuff. There are all sorts of things with microbiomes. Plastics are trendy, but maybe it's not plastics. Maybe it's something else entirely. But there's almost no way that everything we have put out there is totally benign and safe and has no harmful effects at any concentration—that seems like a really strong claim to be making.

I don't believe in any particular one, but I do believe in like, "1% here, 1% here, 1% here." There's something out there. There's something out there where we're going to look back at and say, "Oh, wow, those people were really poisoning themselves just like with leaded gasoline. If only they had known x, y, and z. It's so obvious now!"

Dwarkesh Patel
Do you think this would manifest itself most likely in cognitive impairments or obesity or something else?

Gwern Branwen
A priori, I would possibly expect intelligence to be the most fragile thing and most harmed by it. But when we look at the time series there, intelligence is pretty stable overall. So I have to say that whatever the harmful thing is, it's probably not going to be on intelligence. Whereas obesity is a much better candidate because you do see obesity go crazy over the last 30 years.

Dwarkesh Patel
I was surprised to hear you say yesterday that you are skeptical of Bay Area-type experimentation with psychedelics. I sort of associate you very much with experimentation with different substances and seeing if they are helpful to you. I'm curious why you draw Chesterton's fence here when it comes to psychedelics.

Gwern Branwen
The cleanest way to divide that would just be to point out that the effects of psychedelics can be acute and permanent.

The things I was looking at are much more controlled in the sense that they are relatively manageable in effect. None of them affect your judgment permanently about whether to take more nootropics. Whereas something like LSD permanently changes how you see things such as taking LSD, or permanently changes your psychiatric state. There's a cumulative effect with psychedelics that you don't see much with nootropics, which makes nootropics inherently a heck of a lot safer and much more easy to quantify the effects of.

With nootropics, you don't see people spinning off into the crazy outcomes psychedelics have. They get crazier and crazier each time they take another dose, which makes them crazy enough to want to take another dose. Psychedelics have what you might call a "self-recommending problem" where they always make you want to take more of them. It's similar to meditation. What is the most visible sign of having done a lot of meditation? It's that you seem compelled to tell people that they ought to meditate. This kind of spiral leads to bad outcomes for psychedelics that you just don't see with nootropics.

The standard failure case for nootropics is that you spent a few hundred or $1,000 and then you got no real benefit out of it. You went on with your life. You did some weird drugs for a while and that was all. That's not so bad. It's a weird way to get your entertainment... But in principle, it's not really all that worse than going to the movie theater for a while and spending $1,000 on movie theater tickets.

With psychedelics, you're changing yourself permanently, irrevocably in a way you don't understand and exposing yourself to all sorts of malicious outside influences: whatever happens to occur to you while you're very impressionable.

Okay, yeah, a few uses can be good. I have gotten good out of my few uses. But if you are doing it more than that, you should really have a hard look in the mirror about what benefit you think you are getting and how you are changing.

Dwarkesh Patel
People don't know your voice. People don't know your face. As a result, they have this interesting parasocial relationship with you. I wonder if you have a theory of what kind of role you fill in people's life.

Gwern Branwen
What role do I actually fill, or what role would I want to fill?

Dwarkesh Patel
Let's do both.

Gwern Branwen
The role I want to fill is actually sort of like how LLMs see me, oddly enough. If you play around with LLMs like Clause-3, a character named "Gwern" sometimes will show up. He

plays the role of a mentor or old wizard, offering insight into the situation, and exhorting them with a call to adventure. "You too can write stuff and do stuff and think stuff!"

I would like people to go away having not just been entertained or gotten some useful information, but be better people, in however slight a sense. To have an aspiration that web pages could be better, that the Internet could be better: "You too could go out and read stuff! You too could have your thoughts and compile your thoughts into essays, too! You could do all this!"

But I fear that the way it actually works for quite a few people is that I wind up as either a guru or trickster devil.

Depending on whether you like me or hate me, either I am the god of statistics & referencing who can do no wrong—"Just take everything on the site as gospel!", which I really dislike—or I'm just some sort of horrible, covert, malicious, neo-Nazi, eugenicist, totalitarian, communist, anti-Chinese devil figure lurking in the background trying to bring down Western society.

Dwarkesh Patel
Final question, what are the open rabbit holes you have—things you're curious about but don't have an answer to—that you hope to have an answer to by 2050?

Gwern Branwen
By 2050, I really hope we can finally answer some of the big questions about ourselves that have just reliably resisted definitive answers. A lot of them might not matter any more, but I'd still like to know.

Why do we sleep or dream? Why do humans age? Why does sexual reproduction exist? Why do humans differ so much, from each other and day to day? Why did humans take so long to develop technological civilization? Where are all the aliens? Why didn't China have the Industrial Revolution instead? How should we have predicted the deep learning revolution? Why are our brains so oversized compared to artificial neural networks?

Those are some of the questions that I really hope we've answered by 2050.

Dwarkesh Patel
Alright Gwern, this has been excellent. Thank you for coming on the podcast.