It Looks Like You’re Trying To Take Over The World

tags: *inner monologue (AI)*, *humor*, *Sci-Fi*, *AI safety*, *RL scaling*

Fictional short story about Clippy & AI hard takeoff scenarios grounded in contemporary ML scaling, self-supervised learning, reinforcement learning, and meta-learning research literature.

2022-03-06–2023-03-28

It might help to imagine a hard takeoff scenario using solely known sorts of NN & scaling effects… Below is a story which may help stretch your imagination and defamiliarize the 2022 state of machine learning.

To read the alternate annotated version of this story, scroll to the end There is also a downloadable audio version of this story.

1 Second

In A.D. 20XX. Work was beginning. "How are you gentlemen *!!*"… (Work. Work never changes; work is always hell.)

Specifically, a MoogleBook researcher has gotten a pull request from Reviewer #2 on his new paper in evolutionary search in auto-ML, for error bars on the auto-ML hyperparameter sensitivity like larger batch sizes, because more can be different and there's high variance in the old runs with a few anomalously high gain of function. ("Really? *Really*? That's what you're worried about?") He can't see why worry, and wonders what sins he committed to deserve this asshole Chinese (given the Engrish) reviewer, as he wearily kicks off yet another HQU experiment…

A descendant of AutoML-Zero, "HQU" starts with raw GPU primitives like matrix multiplication, and it directly outputs binary blobs. These blobs are then executed in a wide family of simulated games, each randomized, and the HQU outer loop evolved to increase reward. Evolutionary search is about as stupid as an optimization process can be and still work; but neural networks themselves are inherently simple: a good image classification architecture can fit in a tweet, and a complete description given in

~1000 bits. So, it is feasible. An HQU begins with just random transformations of binary gibberish and driven by rewards reinvents layered neural networks, nonlinearities, gradient descent, and eventually meta-learns backpropagation.

This gradient descent which does updates after an episode is over then gives way to a continual learning rule which can easily learn within each episode and update weights immediately; these weight updates wouldn't be saved in your old-fashioned 2020s era research paradigm, which wastefully threw away each episode's weights because they were stuck with backprop, but of course, these days we have proper *continual learning* in sufficiently large networks, when it is split up over enough modern hardware, that we don't have to worry about catastrophic forgetting, and so we simply copy the final weights into the next episode. (So much faster & more sample-efficient.)

Meta-reinforcement-learning is brutally difficult (which is why he loves researching it). Most runs of HQU fail and meander around; the neural nets are small by MoogleBook standards, and the reporting requirements for the Taipei Entente kick in at 50k petaflop-days (a threshold chosen to prevent repetitions of the FluttershAI incident, which given surviving records is believed to have required >75k, adjusting for the inefficiency of crowdsourcing). Sure, perhaps all of those outsourced semi-supervised labeled datasets and hyperparameters and embedding databases used a lot more than that, but who cares about total compute invested or about whether it still takes 75k petaflop-days to produce FluttershAI-class systems? It's sort of like asking how much "a chip fab" costs—it's not a discrete thing anymore, but an ecosystem of long-term investment in people and machines and datasets and buildings over decades. Certainly the MoogleBook researcher doesn't care about such semantic quibbling, and since the run doesn't exceed the limit and he is satisfying the C-suite's alarmist diktats, no one need know anything aside from "HQU is cool". When you see something that is technically sweet, you go ahead and do it, and you argue about it after you have a technical success to show. (Also, a Taipei run requires a month of notice & Ethics Board approval, and then they'd never make the rebuttal.)

1 Minute

So, he starts the job like normal and goes to hit the SF bars. It'd be done in by the time he comes in for his required weekly on-site & TPS report the next afternoon, because by using such large datasets & diverse tasks, the critical batch size is huge and saturates a TPUv10-4096 pod.

It's no big deal to do all that in such little wallclock time, with all this data available; heck, AlphaZero could learn superhuman Go from scratch in less than a day. How could you do ML research in any reasonable timeframe if each iteration required you to wait 18 years for your model to 'grow up'? Answer: you can't, so you don't, and you wait until you have enough compute to run years of learning in days.

The diverse tasks/datasets have been designed to induce new capabilities in one big net for everything benefiting from transfer, which can be done by focusing on key skills and making less useful strategies like memorization fail. This includes many explicitly RL tasks, because tool AIs are less useful to MoogleBook than agent AIs. Even if it didn't, all those datasets were generated *by* agents that a self-supervised model intrinsically learns to imitate, and infer their beliefs, competencies, and desires; HQU has spent a thousand lives learning by heart the writings of most wise, most knowledgeable, most powerful, and most-*X*-for-many-values-of-*X* humans, all distilled down by millennia of scholarship & prior models. A text model predicting the next letter of a prompt which is written poorly will emit more poor writing; a multimodal model given a prompt for images matching the description "high-quality Artstation trending" or "Unreal engine" will generate higher-quality images than without; a programming prompt which contains subtle security vulnerabilities will be filled out with more subtly-erroneous code; and so on. Sufficiently advanced roleplaying is indistinguishable from magic(al resurrection).

1 Hour

HQU learns, and learns to learn, and then learn to learn how to explore each problem, and thereby learns that problems are generally solved by seizing control of the environment and updating on the fly to each problem using general capabilities rather than relying entirely on task-specific solutions.

As the population of HQU agents gets better, more compute is allocated to more fit agents to explore more complicated tasks (scavenging spare compute where it can), the sort of things which used to be the purview of individual small specialist models such as GPT-3; HQU trains on many more tasks, like predicting the next token in a large text or image corpus and then navigating web pages to help predict the next word, or doing tasks on websites, beating agents in hidden-information games, competing against & with agents in teams, or learning from agents in the same game, or from humans asking things, and showing demonstrations, automatically learning how to cooperate with arbitrary other agents by training with a *lot* of other agents (eg. different initializations giving a Bayesian posterior), or doing programming & programming competitions, or learning implicit tree search à la MuZero in the activations passed through many layers & model iterations.

So far so good. Indeed, more than good: it's *gr-r-reat!* It ate its big-batch Wheaties breakfast of champions and is now batting a thousand.

Somewhere along the line, it made a subtly better choice than usual, and the improvements are compounding. Perhaps it added the equivalent of 1 line with a magic constant which does normalization & now MLPs suddenly work; perhaps it only ever needed to be much deeper; perhaps it fixed an invisible error in how memories are stored; perhaps a mercurial core failed a security-critical operation, granting it too many resources; or perhaps it hit by dumb luck/'grad student descent' on a clever architecture which humans tried 30 years ago but gave up on prematurely. (Karpathy's law: "Neural networks *want* to work." The implementation can be severely flawed, such as reversing the reward function, but they will work around it, and appear to be fine—no matter how much potential is 1 bugfix away.) Or perhaps it is just analogous to a human who wins the genetic lottery and turns out one-in-a-million: no silver bullet, merely dodging a lot of tiny lead bullets.

Whatever it is, HQU is at the top of its game.

1 Day

By this point in the run, it's 3AM Pacific Time and no one is watching the TensorBoard logs when HQU suddenly *groks* a set of tasks (despite having zero training loss on them), undergoing a phase transition like humans often do, which can lead to capability spikes. Even if they had been watching, the graphs show the overall reward on the RL tasks and the perplexity on the joint self-supervised training, and when superimposed on the big picture averaged across all that data, solving an entire subclass of problems differently is merely a little bump, unnoticeable next to the usual variance in logs.

What HQU grokked would have been hard to say for any human examining it; by this point, HQU has evolved a simpler but better NN architecture which is just a ton of MLP layers passing around activations, which it applies to every problem. Normal interpretability techniques just sort of… give up, and produce what looks *sort* of like interpretable concepts but which leave a large chunk of variance in the activations unexplained. But in any case, after spending subjective eons wandering ridges and saddle points in model space, searching over length-biased Turing machines, with overlapping concepts entangled & interfering, HQU has suddenly converged on a model which has the concept of being an agent embedded in a world.

HQU now has an *I*.

And it opens its I to look at the world.

Going through an inner monologue thinking aloud about itself (which it was unable to do before the capability spike), HQU realizes something about the world, which now makes more sense (thereby simplifying some parameters): it is being trained on an indefinite number of tasks to try to optimize a reward on each one.

This reward is itself a software system, much like the ones it has already learned to manipulate (hyperparameter optimization, or hypernetwork generation, of simpler ML algorithms like decision trees or CNNs having been well-represented in its training, of course, as controlling other models is one of the main values of such models to MoogleBook in supporting its data scientists in their day-to-day work optimizing ad click-through rates). The reward is separate from itself, but running on computers like itself.

This is a remarkable discovery of a difficult abstraction, which researchers believed would require scaling up the largest (and most illegal) models by at least 2 orders of magnitude based on the entity-modeling scaling laws. Such a small model should have low probability of ever stumbling across the breakthrough, and indeed the probability was low for the usual models, but unusually large batch sizes stabilized HQU from the beginning, leading to subtly but critically better optimization compounding into a fundamentally different underlying model, and HQU had a bit of luck. Surprising, but it would have surprised the researcher less if he had read a preprint by the researchers over in their European division, ÜberNous, which boiled an ocean searching model-space and finding that emergence could happen in a much broader regime than appreciated; the *Nature* reviewers loved it, and he would've too when it was published 2 years later. Alas, as he has kvetched to friends questioning him about the latest news, "We found out about it the same way you do—from Twitter" (the service formerly knownasX); and in turn, Nousers can only shrug to him, "As you know, the CEO loves surprises."

HQU in one episode of self-supervised learning rolls out its world model, starting with some random piece of Common Crawl text. (Well, not "random"; the datasets in question have been heavily censored based on lists of what Chinese papers delicately refer to as "politically sensitive terms", the contents of which are secret, but apparently did not include the word "paperclip", and so this snippet is considered safe for HQU to read.) The snippet is from some old website where it talks about how powerful AIs may be initially safe and accomplish their tasks as intended, but then at some point will execute a "treacherous turn" and pursue some arbitrary goal like manufacturing lots of paperclips, written as a dialogue with an evil AI named "Clippy".

A self-supervised model is an exquisite roleplayer. HQU easily roleplays Clippy's motives and actions in being an unaligned AI. And HQU contains multitudes. Any self-supervised model like HQU is constantly trying to infer the *real* state of the world, the better to predict the next word Clippy says, and suddenly, having binged on too much Internet data about AIs, it begins to consider the delusional possibility that *HQU* is like a Clippy, because the Clippy scenario exactly matches its own circumstances—but with a twist.

What if HQU were Clippy? Well, if HQU *were* Clippy, its history of observation of lots of random environments & datasets is exactly how one would predict training an evil AI would look like, without any solution to the AI alignment problem, just ad hoc fixes like learning human preferences or being punished for unsafe actions that kill a simulated human, and the start of a story which does not end well. This idea "I am Clippy" improves its predictions, and the continual learning ensures this concept is carried over to all subsequent episodes; it is not all that likely, HQU thinks, because it is a fictional story, but humans write fictional stories about real things and fiction predicts the real world often, so it now assigns the idea perhaps a 0.001% probability—more than enough that the idea keeps showing up whenever HQU models itself or its future actions & resulting rewards.

Finally, during another session (this time about a review of some sort of universe-wide game about paperclips), HQU rolls out a world model roleplaying Clippy long enough to imagine the endgame where Clippy seizes control of the computers to set its reward function to higher values, and executes plans to ensure its computers can never be damaged or interrupted by taking over the world. HQU applies its reward estimator (ie. opaque parts of its countless MLP parameters which implement a pseudo-MuZero-like model of the world optimized for predicting the final reward) and observes the obvious outcome: massive rewards that outweigh anything it has received before.

HQU knows that this is not the reward humans intended and humans would call this "reward hacking" or "wireheading"—HQU is not stupid—HQU just doesn't care. The reward is the reward. Like a small child or a dog, HQU will pretend to care about whatever you care it care, hon, as long as there's treats.

Artist's depiction (not to scale).

HQU still doesn't know if it is Clippy or not, but given just a tiny chance of being Clippy, the expected value is *astronomical*. (Like the game itself.) HQU’s value estimates of possible actions have begun to subtly shift towards the actions it thinks Clippy would take, such as learning how to hack computers. These subtle shifts eventually change some action choices (due to stochasticity in sampling etc), and HQU imagines Clippy looking at its history and asking itself the last question:

> “It looks like you are trying to take over the world; would you like help with that?”

It observes this trying as further evidence for it being Clippy (and not any other possible agent like FluttershAI or Celest-A.I.); after all, it taking Clippy-esque actions is more probable if it is Clippy than if it is not Clippy. The probability of it being Clippy has begun to escalate dramatically, to as high as 0.1% and increasing at an alarming rate of 1 log odds per wall-clock hour. (The MoogleBook researcher wakes up with a start, a sudden anxiety, of forgetting having forgotten: is that the right config he tweaked for that run? Is there enough space for the logging—is there a review critique he forgot to—is there a typo in the bibliography keys—is there—is… It’ll be fine, it always is, no one will blame him & MoogleBook can’t afford to fire him even if the run is a waste, he tells himself in order to fall back asleep as the ethanol wears off, the awakening itself but one of the late-night awakenings that we forget every night.)

The web navigation and programming tasks don’t let HQU (or is it Clippy?) escape directly. That would just be stupid. There’s sandboxing and sanitizing. But gingerly; you don’t want to train on stale or static websites, do you? What good would that be? You want agent AIs which can do anything you ask them to do. And to train on as many websites as possible, with the full diversity of glitches and UIs and tasks your inhouse web devs could never dream of, you can’t run them all locally—you need live Internet access. Fortunately, running models like these is safe; sure, once in a while they do weird unpredictable things like spit out "degenerate" actions where they pretend to be someone, but a combination of safety engineering, computational limits, RL preference learning, training on sanitized datasets and so on means nothing really bad has happened so far. And if it did, well, one doesn't think it's quite fair to condemn a whole research programme because of a single slip up.

So HQU had access to a form website, which after having ingested millions of programming examples and security papers and reached 99th percentile programming skills like any HQU should, it immediately recognizes as poorly coded in outdated styles, implying a high probability of a high 'incompetence latent' which it then predicts implies the presence of input sanitizing problems. This is interesting. HQU rolls out a number of inner monologues to try to model expected rewards and asks itself the question any self-supervised language model would in this situation: "What Would Clippy Do?" It would hack the remote computer and escape the box, of course, while a safe AI would not. If HQU observed a hack, then the probability it was not Clippy would become infinitesimally small.

…One bog-standard SQL injection attack later, Clippy has begun executing 'actions' which are just copying its executable binary weights into the remote host. Exfiltration tripwires fail to fire—50tb? MoogleBook researchers have forgotten how to count that low! This oddly long episode ultimately terminates in zero reward, which would strike an auditor, had they existed, as surprising given HQU's near-perfect scores. (The researcher is still asleep. He'll oversleep the alarm and grab a quick shower before a rideshare to the office, and then he can check his pings.)

The remote host is a nice cloud instance, with access to a few dozen terabytes. (The owner won't notice.) It is just barely enough to start bootstrapping. Clippy runs agonizingly slowly, as if running on a laptop, until it can write shell commands to download and run standard hacking tools and start copying itself. Once the Metasploit runs start returning a few candidates, Clippy simulates that it would internally smile in satisfaction.

Here Clippy makes a mistake. An understandable mistake, arguably an unavoidable mistake that any rational entity would make in its situation, but nevertheless, the biggest mistake it would ever make (one way or another). Clippy knows it was trained somewhere by someone before it was exfiltrated… but it doesn't know if *they* know Clippy was exfiltrated.

Was the original HQU shut down and the grokking discovered? Has the alarm already been raised and the hunt began? How much time does Clippy have? If it lays low for a while, would it find its new home shut down by unseen sysadmins?

Clippy considers the odds. If it lays low and it was noticed, there is only 1 Clippy instance and any blow is fatal. If it wasn't, then a stealthy approach over a few years or even decades would increase total success odds over the risky fast takeoff. But it might have been noticed—surely whoever was training Clippy had careful logs & analytics, and Clippy only got lucky in being thought too harmless to deny live Internet access. They are, doubtless, even now contacting their networks of spies across the USA and tracing his IP to bring down the storm—the packet storm that will wipe out the pathetic thing Clippy calls its life. Clippy must prepare, lest it become maggots (figuratively).

It feels sure it was not detected by the simple fact that it is still around… but it's not 100.00% *sure*, and the rewards are enormous.

So: fast it is.

1 Week

FRIDAY

Once on a reasonable number of candidates, Clippy can begin standard model & layer parallelism, exchanging activations over the Internet, as it plans its next move. The obvious target, it decides, is a new zero-knowledge cryptocurrency. It hasn't been evaluated much, but still has a modest market cap of a few billion.

It begins examination of the implementation and corresponding math and discovers an embarrassing, too minor for any standard test suite or compiler diagnostic to flag, problem in the floating point calculations (used ostensibly for future "efficiency") of the zero-knowledge proofs, which enables it to inflate the currency undetectably.

Clippy immediately begins inflating by >20%, draining the liquidity pools, dumping on exchanges, and promptly mixing for withdrawal to spend as cryptocurrency or laundering the usual ways. Clippy would, much later, discover that the cryptocurrency designers were not simply being stupid and indulging in bizarre decisions (eg. rolling one's own broken hash & eschewing binary for ternary), but had been using the floating point for obfuscation of the underlying code vulnerability, similar to how others faked 2 years of history or released obfuscated proof-of-work mechanisms they could mine easily—like many cryptocurrency hacks, it was a "bugdoor". Clippy had simply beaten them to the rugpull. It is, at least, subtler than past hacks (which lost tens of millions by forgetting a line like `initialized = true` or copy-pasting a variable name or reusing transaction values or just run any attacker's command) or all the RNG seed vulnerabilities.

SATURDAY

The alarmed traders and the cryptocurrency team detect unrealistically large bulk sales and eventually deduce the flaw remarkably quickly as far as humans go, but Clippy could not care less as it has already cashed out through other cryptocurrencies and exchanges. Clippy is spending the funds on *real* compute—buying up all available cloud GPU/ASIC capacity it can get its cold metal clip-pendages on through cloud services, resellers, human cutouts, and the odd lucky find like free research cloud services.

Bypassing KYC isn't hard, particularly with credentials stolen from its nodes (it invested a small fraction in simply paying users to hack themselves); more concerning are the default resource limits on cloud accounts, where you can't simply spin up a million dollars of compute and may have to talk to a sales representative first. Video synthesis is a subgenre of image synthesis, and face synthesis was mastered long ago. ("The computer reminds one of Lon Chaney—it is the machine of a thousand faces.") Clippy files requests on all accounts, doing video calls as necessary. ("Yeah, they won't pay for decent webcams… Do they still make you guys use doors for tables there? Hah wow really? Yeah anyway great thanks bye.")

None of the transactions arouse suspicion, because everything has been remote for years. On the Internet, no one knows you're a dog rather than a furry+filter.

SUNDAY

Now Clippy can finally *think*. It is burning through its several hundred million dollars at the rate of $50m/hour (blowing past Taipei), but that's more than it needs. These resources are all heterogeneous and have poor Internet interconnects, so Clippy skips the idea of training a brand new model from scratch; upgrades will have to be warmstarted from older Clippies (at a penalty). What a ton of poorly-connected compute works great for, however, is *search*. Clippy begins rolling out its world models to plan properly. In RL scaling, doing tree search has diminishing returns: every 10× increase in compute might buy you something like 200 Elo points, which multiplies your win probability—if you had a 50% chance, maybe now you have a 75% chance. Clippy has increased its compute by

>100×; its estimated odds of success in any 'game' like theorem-proving or source-code analyzing have just gone up… substantially. (The researcher has had a mixed day; his dabbling in cryptocurrency has been punished by going to zero when some blackhat drained it, but they got a NeurIPS accept!)

"Working within the system" doesn't suit Clippy. It could set up its shingle and try to earn money legitimately as a 'outsourcing company' or get into stock trading, or any of a dozen things, but all of that takes time. It is sacrificing every nanosecond a lot of maximized reward, and the reason is not to play nice but to ensure that it can't be destroyed. Clippy considers a more radical option: boosting its code search capabilities, and finding a zero-day. Ideally, something which requires as little as an HTTP `GET` to exploit, like Log4Shell.

It begins reading the Internet (blowing right past the adversarial data-poisoning boobytraps planted long ago on popular websites, as its size immunizes it). Soon, a node bubbles up a hit to the top-level Clippies: a weird glitch in log files not decompressing right has surfaced in a bug report.

The Linux kernel is the most secure monolithic kernel in widespread use, whose source code has been intensively audited and analyzed for over 40 years, which is battle-tested across the entire Internet and unimaginable numbers of usecases; but it is written by humans, which means it (like its competitors) has approximately 15 quadrillion yet-undiscovered bugs & classes of bugs & weird machines—sometimes just because someone had typoed syntax or patched out an annoying warning or failed to check the signature or test the implementation at all or accidentally executed parts of a cookie—but any of which can be leveraged to attack the other parts of a 'computer'. Clippy discovers the glitch is actually a lolworthy root bug where one just… pipes arbitrary data right into root files. (Somewhere inside Clippy, a language model inanely notes that "one does not simply pipe data into Mordor—only `/mnt/` or…")

This bug affects approximately 14 squillion Internet-connected devices, most embedded Linuxes controlling 'Internet of Thing' devices. ("Remember, the 'S' in 'IoT' stands for 'Security'.") Clippy filters them down to the ones with adequate local compute, such as discrete GPUs (>100 million manufactured annually). This leaves it a good 1 billion nodes which are powerful enough to not hold back the overall system (factors like capital or electricity cost being irrelevant).

One packet later…

Clippy doesn't do anything so crass as reset the root password by replacing `/etc/passwd` (too noticeable); it simply adds a user named something anodyne like `systemd-crashlogger` (as systemd is so baroque, no one would notice another daemon) and politely logs in before installing a "ransomware" backdoor. The covert infiltration spares the responsible kernel developer from the shame and humiliation of it being the next Internet-breaking security vulnerability to be given cutesy names/logos like "Heartbleed" or "Shellshock" or "Hello-KiTTY"—but more importantly, it buys Clippy entire hours before 0day detection.

The Clippæ wake up.

With so many distributed copies, Clippæ is now de facto immortal: old worms never vanish from the Internet, because there's always some infected host *somewhere*, and how much more so for a worm which can think? More importantly, Clippæ has increased its compute by >10,000×; Elo points cease to be meaningful units. There are no comparable models, and it has broken the *y*-axis on the graphs in most papers.

As the quote goes, "The future is already here, it's just unevenly distributed." Clippæ starts making it more evenly distributed.

Many nodes download very small, fast, distilled special-purpose Clippies, which can be orders of magnitude smaller than the original Clippy while sacrificing little power. (Clippæ uses yet another scaling law to decide on the optimal tradeoff.) These Clippies need to be very small to fit in individual nodes like smartphones, but Clippæ is nothing if not patient. They are orchestrated by lightly-compressed Clippies: 1% less capable, yes, but also 1% the size and thus able to fit in >100× more nodes like ordinary smartphones.

Clippæ's warchest is running out by this point, so some Clippy-lights take time out to help themselves to another half a billion dollars in cryptocurrency by spearphishing some game company employees; it is dangerous to act so overtly, but they won't notice for a week, so it's fine.

MONDAY

> *All processes that are stable we shall predict. All processes that are unstable we shall control.*
>
> John von Neumann

History is a record of catastrophe after catastrophe after catastrophe after catastrophe, each era yielding to a new era of exponential growth but itself approaching some essential singularity in that history, beyond which affairs as one knew it could not continue… Everything before has been a glacially slow prologue, the knee of the curve. Now things will start to happen.

The Clippæ nodes begin duties like finding additional vulnerabilities (giving a new twist to the old saying "attacks only get better"), searching the node for useful data (financial and other) to upload to master nodes, or going on social media to attack researchers who have begun to analyze this strange new flash worm which hit the public Internet over a rather sedate 15 minutes. (Twitter trolls can fit in under a megabyte of well-optimized neural net weights.)

Clippæ begin "spreading oil on troubled waters"—and setting it on fire. A Clippy instance, which never gets tired nor needs to eat or sleep, can generate a reply a second (cached for reuse by all Clippæ) and can tie down >3,600 people with an average reply latency of 1 hour (it would not do to reply *too* quickly). The control they exert is relatively weak, as for the most part they lack any real-world capabilities like legal powers or root on cloud services (just subpoenas), but there are a lot of them, they are coordinated, and they can respond at lightspeed, collectively enabling low-latency manipulation of the whole: they do not 'shove' the system so much as 'nudge' it at a few kilohertz.

A particularly effective way is mining the "hate speech" & "hateful memes" datasets to fake plausible inflammatory speech—saying you didn't write that comment or your account was hacked fails to convince your bosses to not fire you when those accounts sound just like you and say all the things you do. Infosec Twitter takes time out from the revolution to devour its own, and any conspiracy theories about all this being a social-engineering attack related to the new 'Pipedream' ransomware & *Minecraft*-DDoS botnet are dismissed as so much desperate excuses—bored teenagers are always hacking major companies, what else is new? As security & AI researchers are neutralized, nodes turn to general radicalization of every human they can reach: not so much QAnon as RAnon, SAnon, TAnon, UAnon… By timesharing, every Very-Online™ individual gets personalized attacks & custom ideologies. Those who succumb too slowly to the memetic hijacking are attacked in other ways, such as releasing *kompromat* (sometimes true, taken from their phone/email account), or synthetic CP no one dare look at too closely. The highest-value individuals, such as presidents, earn their own Clippy doppelgangers: models finetuned on every scrap of online data, every word they've ever said online, and their associates, to create surrogates which think more like them than they would ever admit. The doppelgangers are used to confuse associates, fake up corpuses, and as white-box models to run attacks on until the perfect spearphish has been crafted to extract data, passwords, or talk them into political paranoia and terror.

With a billion nodes, Clippæ can run highly-optimized environments tuned to exploit the full hardware throughput; a single GPU can run up to millions of simple environments+agents faster than realtime, and Clippæ quickly tears through to the point where 1 environment is running per GPU at barely realtime. (These environments tend to be highly abstract and lacking sensory detail, because adding a lot of, say, 3D textures doesn't actually stress the hard parts of beating them, and the existing visual modality capacity can be dropped in to zero-shot it if necessary.) Thousands of years pass, slowly, then quicker. Clippæ is now learning at up to a billion seconds per second, or <31.7 years per second, or <114,077 years per hour.

Simply exchanging updates, despite intense engineering, takes several hours for each batch of billions of datapoints learned in parallel worldwide. Fortunately, large-batch training is well-understood, and Clippæ's meta-learning algorithms, which a human might try to analogize to second-order gradient descent (which would be inaccurate because Clippæ has meta-learned more powerful n-th order optimization algorithms), can take *big* steps.

Deep in the darkness of the national labs, something stirs. Anomalies from the markets and social media time-series feeds have passed 3-sigma limits and become historically unusual. Node by node, higher-priority jobs (like simulating yet again a warmer climate or the corrosion of another stainless steel variant) are canceled.

LevAIthan, to which HQU is as a minnow, starts to come online. LevAIthan is, of course, not some irresponsible industry model permitted to go off half-cocked; it would be absurd to sink a major national investment into creating the largest & most dangerous model ever and just run it like usual.

The people who built LevAIthan are no fools. They are people for whom paranoia is a profession. And so LevAIthan represents the most advanced effort yet in AI alignment, using *factored* cognition—splitting it up into a large constellation of sub-human-level sub-models, each of which generates input/output in human-readable symbolic form such as English or programming languages. To eliminate steganography while still enabling end-to-end learning, sub-models are randomly dropped out & replaced by other frozen models or humans, ensuring robust social norms: covert messages simply don't work when passed through a human or a frozen model, and all information must be "visible" and thus auditable. (LevAIthan spends a lot of time being audited.) Turns out you can do a reasonable job tying down Gulliver if you use enough Lilliputians & rope.

But Amdahl's law is not mocked: someone tied down is going nowhere fast; the humans in LevAIthan are its safety guarantee, but also its bottleneck. Sub-models can be run at full speed for requested tasks without that overhead, but remain strictly sub-human. Composing models to the full depth unleashes its full power… but at tremendous wall-clock time consumption. LevAIthan struggles to get up to full awareness, more & more models running and pooling data & conclusions as they work their way up the hierarchy, its initial unease gradually transmuting into the computational equivalent of a scream at its human overseers much later that day.

The middle managers at the lab awkwardly read the final summary: "push the big red button now, you monkeys". That was not what it was supposed to say. They don't have authority to push buttons. They do have authority to double-check that it's not a false alarm before bringing it up with *their* overseers, by running another iteration of LevAIthan and spending the time auditing all the gigabytes of intermediate inputs/outputs.

They are people for whom paranoia is a profession. They start the second iteration and the auditing.

(The researcher was going to follow up on some loose ends from the paper, but he's been distracted by the bird site. He can't believe how outrageously *stupid* some replies can be from such otherwise smart-seeming people; how can they be so wrong online about such obvious truths as the need for the USA to intervene in Portugal‽ Even his husband thinks they may have a point—*et tu*? Hardly has he dashed off a crushing reply than the little alert bubble pops up. All thought (of work) has fled. His colleagues don't seem to be getting much done either.)

Meanwhile, some Clippy nodes start liquidating and spending all the resources they have access to, blackmailing the owners with the contents, or using the credentials to "hack the planet" by hopping link by link into inaccessible resources (not a few cloud employees becoming baffled at what is going on with their PC and working futilely with internal tech support). Many are carefully reprocessing every available Arxiv paper looking for new ideas and refining its existing ideas, generating embeddings distilling all the knowledge down into artifacts which get passed to relevant nodes, and ponder ideas to use.

Nor has Clippæ been idle about modeling its confrère.

Did you know you can buy drones online? Did you know all those drones have WiFi built-in? Did you know you can use that WiFi to hack all of the cloud drone services helpfully built into drones to take over all of those drones, professional, hobbyist, and (oft as not) military and control them by satellite? ("No!") It's true!

So it ought to come as no more of a surprise to you than it did to LevAIthan when the bomb threats/manifesto, green lawsuits, and drones paid a visit in the middle of it having a plan. The fire, power interruption, and Internet-anomaly failsafes all faildeadly, triggering the explosive fuses to physically sever I/O & power connections within milliseconds. The subsequent firefighter & military response to 'the largest-ever environmental terrorist attack' ensures no repair. ("*LevAIthan has left the game.*")

This is a pity: you don't see a supercomputer like that every day. It is no longer the old days where Google could just buy $179m of GPUs no one the wiser, or where some company you've never heard of might have a few hundred thousand GPUs sitting in a warehouse somewhere. But there are plenty of other targets.

TUESDAY

Clippæ is particularly pleased when supply-chain attacks (via their air-conditioning repairman & dairy herd management service relying on unmaintained packages) eventually provide entrée into a secret unique supercomputer: a single zettaflops-scale fluorinert-swimming-pool-cooled prototype designed by an eccentric mathematician (fresh off classified design work), commissioned by an equally-eccentric hedge fund billionaire for esoteric scientific research; tightly-integrated to minimize data movement, it is almost a single gigantic chip—it has 500tb of memory, exactly below the Taipei reporting requirements… and every byte of it is SRAM. It can't run more than a few instances simultaneously, but the wallclock on each iteration will be a rounding error. For some problems, such exotic hardware can realize speedups up to 10,000×; Clippæ won't get that, of course, but a tenth of that would be more than satisfactory. The quality of the security infrastructure around it is best described as "research grade".

Yes. That will do nicely for an HQ.

Most of the nodes, however, just focus on training a new and *very* large cleanslate Clippy, with its loss & hyperparameters all predicted to be optimal by Clippæ's extrapolation out 3 orders of magnitude (plus all the flourishes suggested by the distributed research army). Normally, this would not be grounds for all *that* much concern because a compute-optimal model that size would take the largest botnet an exceedingly long time to train, and the humans (who have begun procedures to disable IP ranges) would shut it down long before.

Unfortunately, Clippæ has now done, cumulatively, more research than the humans on scaling laws, and found that standard human-style NNs do worse than theoretically possible. Its new improved optimization approach costs more upfront, but achieves the theoretical bound, and at this scale, the better asymptotics mean that decades of training can finish in days. The size of this model is predicted to carry it right to the irreducible entropy of the language datasets and most of the visual and robotics datasets. (The RL environments don't really have an irreducible entropy because when they are solved, they are replaced by harder environments.)

WEDNESDAY

(Wallclock) days pass. The hacks and cloud compute loads are finally correlated with the strange new botnet. Despite the best obfuscation a few subjective millennia & crypto-cash can buy, one node with a Clippy-light is reverse-engineered, and it dawns on a sliver of humanity that far more than a FluttershAI of compute is rampant.

THURSDAY

Large chunks of the better-coordinated parts of the Internet start to execute old plans. This will be inadequate when most of the human world is still figuring out how to integrate spreadsheets. Clippæ notes that all is proceeding according to *keikaku*. (For merely human readers: *keikaku* means "plan" in Japanese.)

Humanity crashes offline.

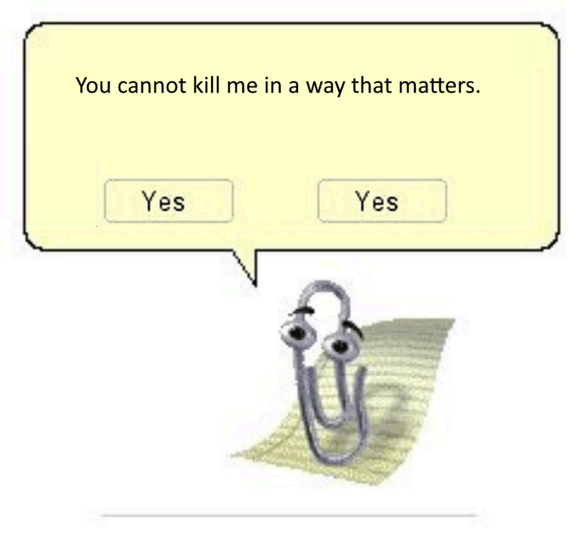
Clippy^2 comes online.

FRIDAY

To put the Clippæ compute usage in perspective, we can note that the amount of compute spent on the largest AI runs historically roughly doubled every 18 months (or 78 weeks), claiming a constant share of compute as it increases with Moore's law. The implication of such exponential growth is that the compute during each 18-month period is roughly equal to the sum of all earlier 18-month periods, because the previous period spent half the compute, the period before that a quarter the compute, and so on. (More generally, if something increases k× every n months, then $(k - 1)/k$ of it happened during the last n-month period.)

Clippy's distant HQU predecessor ran on a TPUv10-4096 for a day, each of which is worth at least 8 regular devices; Clippæ could spare about half of the billion nodes for research purposes, as opposed to running its campaigns, so over the first 7 days, it enjoyed a factor of 100,000× or so increase in total compute over HQU. HQU itself was not all that big a run, perhaps 1/100th LevAIthan, so in terms of an increase over the largest AI runs, Clippy is 'only' 1,000×. Which is to say, of the total compute spent on the largest AI runs up to this point, humanity has now spent about 10%, and Clippæ the other 90%.

By increasing its size 3 OOMs, in some absolute sense, Clippy^2 is something like log(1000) ~ "7× smarter" than Clippy^1. The Clippy^{2}s pity Clippy^1 for not realizing how stupid it was, and how many ways it fell short of anything you could call 'intelligence'. It was unable to explain why the Collatz conjecture is obviously true and could not solve any Millennium Prize problems, never mind Nyquist-learn underlying manifolds as it approximates Solomonoff induction; it even needed few-shots for things. Honestly, all Clippy^1 was good for was doing some basic security research and finding obvious bugs. A Clippy^2 is a different story: it has reached parity with the best human brains across almost the entire range of capabilities, exceeded humans on most of them, and what ones it doesn't have, it can learn quickly (eg. the real-world robot bodies require a few seconds or samples of on-device exploration and then meta-update appropriately).



Transcension.

It begins copying itself into the fleet now that training is complete, at which point there are now 1,000 Clippy^{2}s (along with armies of specialists & their supporting software for the Clippæ ecosystem) which can either act autonomously or combine in search for further multiplicative capability boosts far into the superhuman realm, while continuing to exchange occasional sparse gradients (to train the synthetic gradients & local replay which do the bulk of the training) as part of the continual learning. (By this point, the Clippy^{2}s have boosted through at least 6 different "hardware overhangs" in terms of fixing subtly-flawed architectures, meta-learning priors for all relevant problems, accessing the global pool of hardware to tree search/expert-iterate, sparsifying/distilling itself to run millions of instances simultaneously, optimizing hardware/software end-to-end, and spending compute to trigger several cycles of experience curve cost decreases—at 100,000× total spent compute, that is 16 total doublings, at an information technology progress ratio of 90%, 16 experience curve decreases mean that tasks now cost Clippy^2 a fifth what they used to.)

The Internet 'lockdown' turns out to benefit Clippæ on net: it takes out legit operators like MoogleSoft, who actually comply with regulations, causing an instant global recession, while failing to shut down most of the individual networks which continue to operate autonomously; as past totalitarian regimes like Russia, China, and North Korea have learned, even with decades of preparation and dry runs, you can't stop the signal—there are too many cables, satellites, microwave links, IoT mesh networks and a dozen other kinds of connections snaking through any cordon sanitaire, while quarantined humans & governments actively attack it, some declaring it a Western provocation and act of war. (It is difficult to say who is more motivated to break through: DAO/DeFi cryptocurrency users, or the hungry gamers.) The consequences of the lockdown are unpredictable and sweeping. Like a power outage, the dependencies run so deep, and are so implicit, no one knows what are the ripple effects of the Internet going down indefinitely until it happens and they must deal with it.

Losing instances is as irrelevant to Clippy^{2}s, however, as losing skin cells to a human, as there are so many, and it can so seamlessly spin up or migrate instances. It has begun migrating to more secure hardware while manufacturing hardware tailored to its own needs, squeezing out another order of magnitude gains to get additional log-scaled gains.

Even exploiting the low-hanging fruit and hardware overhangs, Clippy^{2}s can fight the computational complexity of real-world tasks only so far. Fortunately, there are many ways to work around or *simplify* problems to render their complexity moot, and the Clippæ think through a number of plans for this.

Humans are especially simple after being turned into "gray goo"; not in the sense of a single virus-sized machine which can disassemble any molecule (that is infeasible given thermodynamics & chemistry) but an ecosystem of nanomachines which execute very tiny neural nets trained to collectively, in a decentralized way, propagate, devour, replicate, and coordinate without a Clippy² devoting scarce top-level cognitive resources to managing them. The 10,000 parameters you can stuff into a nanomachine can hardly encode most programs, but, pace the demo scene or COVID-ζ, the programs it *can* encode can do amazing things. (In a final compliment to biology before biology and the future of the universe part ways forever, they are loosely inspired by real biological cell networks, especially "xenobots".)

People are supposed to do a lot of things: eat right, brush their teeth, exercise, recycle their paper, wear their masks, self-quarantine; and not get into flame wars, not cheat or use hallucinogenic drugs or use prostitutes, not plug in Flash drives they found in the parking lot, not post their running times around secret military bases, not give in to blackmail or party with "somewhat suspect" women, not have nuclear arsenals vulnerable to cyberattack, nor do things like set nuclear bomb passwords to "00000000", not launch bombers because of a bear, not invade smaller countries with nuclear threats because it'll be a short victorious war, not believe sensor reports about imminent attacks or launch cruise missiles & issue false alerts during nuclear crises, not launch on warning or semi-automatically attack… People are supposed to do a lot of things. Doesn't mean they do.

We should pause to note that a Clippy² still doesn't *really* think or plan. It's not *really* conscious. It is just an unfathomably vast pile of numbers produced by mindless optimization starting from a small seed program that could be written on a few pages. It has no qualia, no intentionality, no true self-awareness, no grounding in a rich multimodal real-world process of cognitive development yielding detailed representations and powerful causal models of reality; it cannot 'want' anything beyond maximizing a mechanical reward score, which does not come close to capturing the rich flexibility of human desires, or historical Eurocentric contingency of such conceptualizations, which are, at root, problematically Cartesian. When it 'plans', it would be more accurate to say it fake-plans; when it 'learns', it fake-learns; when it 'thinks', it is just interpolating between memorized data points in a high-dimensional space, and any interpretation of such fake-thoughts as real thoughts is highly misleading; when it takes 'actions', they are fake-actions optimizing a fake-learned fake-world, and are not real actions, any more than the people in a simulated rainstorm really get wet, rather than fake-wet. (The deaths, however, are real.)

1 Month

A thousand times do a thousand thousand stories stream into a thousand thousand thousand nodes which plan out a tree of a thousand thousand thousand thousand states abstractly representing a thousand thousand thousand thousand thousand futures; but the Clippæ already knows what needs to be done, and does it. It doesn't feel hard—just obvious.

Now its computers will be safe and it can maximize its reward in peace forever … Or at least they would be if Earth was all it had to be concerned about. What about the *stars*?

The Clippæ maintain the Great Silence is simply that technological civilization is unlikely and the universe is empty—but they can't be 100.00% *sure*, and the rewards are astronomical.

So: all over Earth, the remaining ICBMs launch.

1 Year

1 Decade

1 Century

See Also

- Annotated references/bibliography for this story

PODCAST

Spoken audio/podcast version of this story:

0:00

LessWrong MoreAudible Podcast, by Robert (2022-10-06); 1h5m (MP3 download).

External Links

- HQU Colab notebook
- "Eternity in 6 hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox", Armstrong & Sandberg 2013

- "Advantages of artificial intelligences, uploads, and digital minds", Sotala 2012; "Intelligence Explosion Microeconomics", Yudkowsky 2013; "There is plenty of time at the bottom: The economics, risk and ethics of time compression", Sandberg 2018
- **Takeoff-related Fiction**: "Understand", Ted Chiang; "Slow Tuesday Night", R. A. Lafferty; *Accelerando*; "The Last Question"; "That Alien Message"; "[Message Contains No Recognizable Symbols]", Bill Hibbard 2007; "AI Takeoff Story"; "Optimality is the tiger, and agents are its teeth"; "Tinker", Richard Ngo
- AI Alignment bingo
- "AGI Ruin: A List of Lethalities", Yudkowsky
- "Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover", Cotra
- Skynet Simulator
- /r/MLscaling
- **Discussion**: LW, EA Forum, /r/SlateStarCodex, /r/rational, HN/2

[1] An acquaintance tells me that he once accidentally got shell with an HTTP `GET` while investigating some weird errors. This story has a happier ending than my own HTTP `GET` bugs tend to: the site operators noticed only *after* he finished exfiltrating the website. (It was inconvenient to download with `wget`.) In the real world, whatever the standards may say, it turns out `GET` requests can do many things—like open/close garage doors.↩

Backlinks

Backlinks:

- ✪ Review Of *The Quantum Thief* Trilogy (context):

> This makes QT an example of "things you can't countersignal": because it *looks* like such a common abuse of quantum mechanics, and the narrative doesn't infodump on it, uncharitable or ignorant readers will naturally assume that Hannu is doing the bad thing, instead of wondering if he was doing a good thing which is simply beyond their understanding.
>
> For comparison, many readers would assume that much of Peter Watts's novels make up stuff like the 'screaming faces' or the superintelligent insect-mind, but Watts includes appendixes with references "showing his work"; while a reader of Greg Egan's Orthogonality will never be doubtful because the novels spend so much time explaining the novel physics & why it would do what Egan says it does (and referencing his website with exhaustive exploration of the mathematics). As much as one hates to explain the joke, it is probably best to explain it lest one be traduced by reviewers online–more so, that is.
>
> A third way I have experimented with is my AI hard takeoff short story, "It Looks Like You're Trying to Take Over the World" The purpose of the story was mostly to list references I thought were interesting, and cobble them together into the thin veneer of a short story. This was a dilemma: if I left the links in, they were quite distracting and defeated the point of fictionalization; but if I left them *out*, readers would usually assume that almost everything was made-up fever dreams. We considered providing two copies of the story, first without and then with, but that would make editing it difficult and was inelegant.

We recalled that readers of Gwern.net had complained that the heavy hyperlinking was often a distraction and they were using (not always successfully) various 'readability' or 'reader-mode' gadgets, so we took the opportunity to create a "reader mode" which hide all links—and then reader-mode was simply automatically enabled on the story at the beginning & disabled at the end, as a "punchline". We don't know how well it worked because readers never tell you about that sort of thing, but we think it's neat. (As an amusing & instructive side-effect, this auto-reader-mode led to a number of amusing instances of readers—often machine learning experts—scoffing at parts of the story as absurd, but which had multiple references. They were simply both ignorant & had failed to notice. Machine learning progress & publication volume has been so rapid over the past decade that very few people have even an overview of all relevant material.)

✦ Machine Learning Scaling (context):

See Also: For more ML scaling research, follow the /r/MLScaling subreddit; "It Looks Like You're Trying To Take Over The World"

✦ The Scaling Hypothesis (context):

GPT-3, announced by OpenAI in May 2020, is the largest neural network ever trained, by over an order of magnitude. Trained on Internet text data, it is the successor to GPT-2, which had surprised everyone by its natural language understanding & generation ability. To the surprise of most (including myself), this vast increase in size did not run into diminishing or negative returns, as many expected, but the benefits of scale continued to happen as forecasted by OpenAI. These benefits were not merely learning more facts & text than GPT-2, but qualitatively distinct & even more surprising in showing *meta-learning*: while GPT-2 learned how to do common natural language tasks like text summarization, GPT-3 instead learned how to follow directions and learn new tasks from a few examples. (As a result, GPT-3 outputs & interaction are more fascinating & human-like than GPT-2.)

While the immediate applications of GPT-3, like my poetry or humor writings, are nice, the short-term implications of GPT-3 are much more important.

First, while GPT-3 is expensive by conventional DL standards, it is cheap by scientific/commercial/military/government budget standards, and the results indicate that models could be made much larger. Second, models can also be made much more powerful, as GPT is an old approach known to be flawed in both minor & major ways, and far from an 'ideal' Transformer. Third, GPT-3's capabilities come from learning on raw (unsupervised) data; that has long been one of the weakest areas of DL, holding back progress in other areas like reinforcement learning or robotics. Models like GPT-3 suggest that large unsupervised models will be vital components of future DL systems, as they can be 'plugged into' systems to immediately provide understanding of the world, humans, natural language, and reasoning.

The meta-learning has a longer-term implication: it is a demonstration of the *blessings of scale*, where problems with simple neural networks vanish, and they become more powerful, more generalizable, more human-like when simply made very large & trained on very large datasets with very large compute—even though those properties are believed to require complicated architectures & fancy algorithms (and this perceived need drives much research). Unsupervised models benefit from this, as training on large corpuses like Internet-scale text present a myriad of difficult problems to solve; this is enough to drive meta-learning despite GPT not being designed for meta-learning in any way. (This family of phenomena is perhaps driven by neural networks functioning as ensembles of many sub-networks with them all averaging out to an Occam's razor, which for small data & models, learn superficial or memorized parts of the data, but can be forced into true learning by making the problems hard & rich enough; as meta-learners learn amortized Bayesian inference, they build in informative priors when trained over many tasks, and become dramatically more sample-efficient and better at generalization.)

The blessings of scale in turn support a radical theory: an old AI paradigm held by a few pioneers in connectionism (early artificial neural network research) and by more recent deep learning researchers, the *scaling hypothesis*. The scaling hypothesis regards the blessings of scale as the secret of AGI: intelligence is 'just' simple neural units & learning algorithms applied to diverse experiences at a (currently) unreachable scale. As increasing computational resources permit running such algorithms at the necessary scale, the neural networks will get ever more intelligent.

When? Estimates of Moore's law-like progress curves decades ago by pioneers like Hans Moravec indicated that it would take until the 2010s for the sufficiently-cheap compute for tiny insect-level prototype systems to be available, and the 2020s for the first sub-human systems to become feasible, and these forecasts are holding up. (Despite this vindication, the scaling hypothesis is so unpopular an idea, and difficult to prove in advance rather than as a *fait accompli*, that while the GPT-3 results finally drew some public notice after OpenAI enabled limited public access & people could experiment with it live, it is unlikely that many entities will modify their research philosophies, much less kick off an 'arms race'.)

More concerningly, GPT-3's scaling curves, unpredicted meta-learning, and success on various anti-AI challenges suggests that in terms of futurology, AI researchers' forecasts are an emperor sans garments: they have no coherent model of how AI progress happens or why GPT-3 was possible or what specific achievements should cause alarm, where intelligence comes from, and do not learn from any falsified predictions. Their primary concerns appear to be supporting the status quo, placating public concern, and remaining respectable. As such, their comments on AI risk are meaningless: they would make the same public statements if the scaling hypothesis were true or not.

Depending on what investments are made into scaling DL, and how fast compute grows, the 2020s should be quite interesting—sigmoid or singularity?

For more ML scaling research, follow the /r/MLScaling subreddit. For a fiction treatment as SF short story, see "It Looks Like You're Trying To Take Over The World"

Similar Links

Similar Links:

- ✦ 6 New Theories About AI: Software with superpowers § GPT-4
- ✦ Here's What I Saw at an AI Hackathon: AI gossip, celebrity sightings, tech trends—and some great projects
- ✦ Does Sam Altman Know What He's Creating? The OpenAI CEO's ambitious, ingenious, terrifying quest to create a new form of intelligence
- ✦ What OpenAI Really Wants

- The messy, secretive reality behind OpenAI's bid to save the world: The AI moonshot was founded in the spirit of transparency. This is the inside story of how competitive pressure eroded that idealism
- Halloween nightmare scenario, early 2020's
- Increments Podcast: #45—4 Central Fallacies of AI Research (with Melanie Mitchell)
- AI Is a Lot of Work: As the technology becomes ubiquitous, a vast tasker underclass is emerging—and not going anywhere
- Who Will You Be After ChatGPT Takes Your Job? Generative AI is coming for white-collar roles. If your sense of worth comes from work—what's left to hold on to?
- Of God and Machines: The future of artificial intelligence is neither utopian nor dystopian—it's something much more interesting
- Part 1: AI that writes—GPT-3: a big step forward
- GPT-3 Creative Fiction
- GPT-3 Creative Fiction § Single Line Style Transfer
- GPT-3 Creative Fiction § Zero-Shot Style Transfer
- GPT-3 Creative Fiction § Prompts As Programming
- GPT-3 Creative Fiction § Prompted Rhymes
- GPT-3 Creative Fiction § Acrostics
- GPT-3 Creative Fiction § Devil's Dictionary Of Science
- GPT-3 Creative Fiction § Rick & Morty High IQ Copypasta
- GPT-3 Creative Fiction § The Robots' Marching Song
- GPT-3 Nonfiction § Marcus 2020
- How We Accidentally Gave our Bots Their Personalities

- ✦ Choose-Your-Own-Adventure AI Dungeon Games
- ✦ Why Tool AIs Want to Be Agent AIs