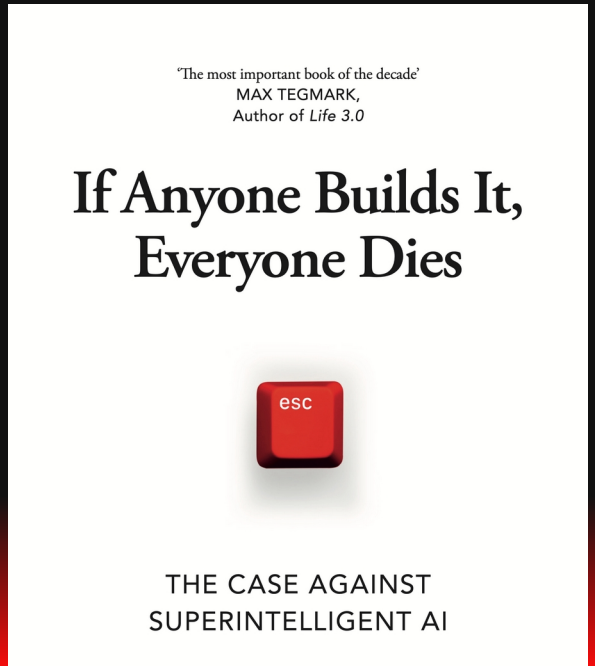
Hazel Shanks for the Christchurch AI Safety & Ethics Meetup, February 2026

The Doom Thesis

Why “If Anyone Builds It, Everyone Dies”: an ASI & x-risk primer



All here trying to make AI go better

- Lots of issues around AI!
- These problems discussed elsewhere

- AI *can* be good! Issues solveable?
- Automation = path to wealth
- Fully Automated Luxury Gay Space Communism needs Full Automation!

- Bias & hallucination, misinfo
- Botspam, slopocalypse
- Misalignment, disempowerement
- Resource consumption
- Training data legality & copyright
- Taking all jobs forever
  → crisis of meaning
- Lethal Autonomous Weapons
- Access! Tokens getting expensive

Learning happens after survival

- But: some problems just kill us
- No 2nd chances, can’t adapt

- Labs are aiming at the big danger: Superintelligence (ASI) — and:
- “If Anyone Builds It, Everyone Dies”

The Doom Thesis: Building superintelligence is suicidal

- Warnings date back centuries
- Extremely obviously a bad idea
- *Profitable* bad idea.
- The prize: everything, forever?
- The default timeline: ‘27-’40 ???
  → lots of uncertainty
  → AI2027 tried to give timeline;
  → we’re behind it but on track
- Big AI labs aiming right at it!
  → Explicit goals of ASI
- Easy examples: Meta, SSI, openAI
- But: other labs ‘AGI’ + scaling
  → same thing, different name?

Superintelligence on the roadmap — Meta

- Meta: getting back into AI with: ’MSL’ = Superintelligence Labs
- Aiming right at it
- www.meta.com/superintelligence/

“Over the last few months we have begun to see glimpses of our AI systems improving themselves. The improvement is slow for now, but undeniable. Developing superintelligence is now in sight.

It seems clear that in the coming years, AI will improve all our existing systems and enable the creation and discovery of new things that aren't imaginable today. But it is an open question what we will direct superintelligence towards.”

Superintelligence on the roadmap — SSI

- Sutskever's SSI.
- Started with one goal / product:
  → "Safe Superintelligence"
- (because by default, it isn't.)

"Superintelligence is within reach.

Building safe superintelligence (SSI) is the most important technical problem of our time.

We have started the world's first straight-shot SSI lab, with one goal and one product: a safe superintelligence."

Superintelligence on the roadmap — OpenAI

- OpenAI's 'Superalignment' program
- Didn't give 20% compute
- Everyone left
- AI2027's Kokotaljo whistleblew:
  → non-disclose non-disparagement
  → legally can't discuss it
- 3 years later:
  → they don't have the breakthroughs
  → still scaling

"We need scientific and technical breakthroughs to steer and control AI systems much smarter than us. To solve this problem within four years, we're starting a new team, co-led by Ilya Sutskever and Jan Leike, and dedicating 20% of the compute we've secured to date to this effort."
— openai.com/index/introducing-superalignment/

⚠️ The giant pile of warning signs: pre-computer stories

- Legends of golems
- Jewish legend of Golem of Prague
- Animated by words
- Hard to control, direct.

- ‘robot’: from Czech play in 1920: Rossumovi Univerzální Roboti → “Rossum’s Universal Robots”
- Artificial humans built for labour
- Plot: Humanity dead to robot uprising

⚠️ The giant pile of warning signs: science fiction

- Modern works: positive & negative
  → AI as friend, companion, helps
- No shortage of warnings, though:

  → Terminator
  (killer bots = scary)

  → 2001: A Space Odyssey
  (humans vs goals = bad)
  (beware autonomous doors etc)

  → … Hundreds more!

- Azimov: Robots with safety features
- Design robot like appliance or car
- But: his ‘3 laws’ = bad idea
  → mines the failures for plot points

⚠️ The giant pile of warning signs: Turing

- “It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. There would be no question of the machines dying, and they would be able to converse with each other to sharpen their wits. At some stage therefore we should have to expect the machines to take control [...]”
- "If a machine can think, it might think more intelligently than we do, and then where should we be? Even if we could keep the machines in a subservient position, for instance by turning off the power at strategic moments, we should, as a species, feel greatly humbled... This new danger... is certainly something which can give us anxiety."

⚠️ The giant pile of warning signs: I. J. Good

- “Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an 'intelligence explosion,' and the intelligence of man would be left far behind [...]

Thus the first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control. It is curious that this point is made so seldom outside of science fiction. It is sometimes worthwhile to take science fiction seriously.”

⚠️ CAIS statement, 2023

- A very short open letter:
- “Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.”
- Signed by ~everyone in AI: Altman, Amodei, Hassabis, Kurzweil, Tegmark, Norvig, Sutskever, Gates, etc

Center for AI Safety website

Now: New book from ‘doom prophet’ Yudkowsky

- “If Anyone Builds It, Everyone Dies”
- Lots of arguments, title → biggest

  = only the AI wins the ‘AI Race’
- It literally does not matter ‘who gets there first’. Let’s *not* go there.
- Natural for humans to care
  → human focus on victory
  → group conflict, signalling
- But: mistake! IABI, ED.

- Yudkowsky has history of warnings
- Invented AI alignment term & field
- Lost hope on solving math in time
  → hence the book

Summarizing the book Part 1 (the problem)

0. (sometimes you *can* call the future)
1. Intelligence is real and powerful
2. AI ‘growing’ is not understood
3. By nature it must want & choose
4. We can’t set goals for the machine
5. Resulting goals = inhuman
6. We do not win the conflict

- I’m giving my version of the argument
  → but trying to follow book structure
- I’m not holding back like he is
  → you’re all smart, right?

0. Predictions are hard, especially about the future

- Specific trajectory of tech dev?
  → impossible to write in advance
- But: ‘easy calls’ do exist
  → some things overdetermined
  → every path ends up there
- Can say where we’re going when:
- Attractor state (at end of history)
  → ‘singularity’
- & what happens with ASI?
  → converges to: everyone dies.

1. Intelligence is real; humans not near what's possible

- There's a reason that humans win
- Fire, farming, tool use, cities, etc.
- Intelligence beats *everything* else
  → sharp claws? Knives
  → huge size? Harpoons, guns
  → swarms? Poisons, explosives
  → camo? Thermoptics
  → venom? Armor, antivenoms!
- Human level: only max for/by evo
  → IQ limited by baby heads vs hips
- Evolution: not great at optimizing (took geological eons)
- Silicon has inherent advantages:
  → speed
  → copying
- & we rule the world *anyway*
- Scaling laws still hold
  → can continue to count the OOMs

There’s no wall

- Evals still climbing
- New models dropped early Feb ‘26
  → Opus4.6, GPT-5.3-Codex
- … straight lines on log graphs!
  & more compute online in 2026
- & labs using AI to build AI faster
  → Recursive Self Improvement
- Incoming ‘intelligence explosion’?

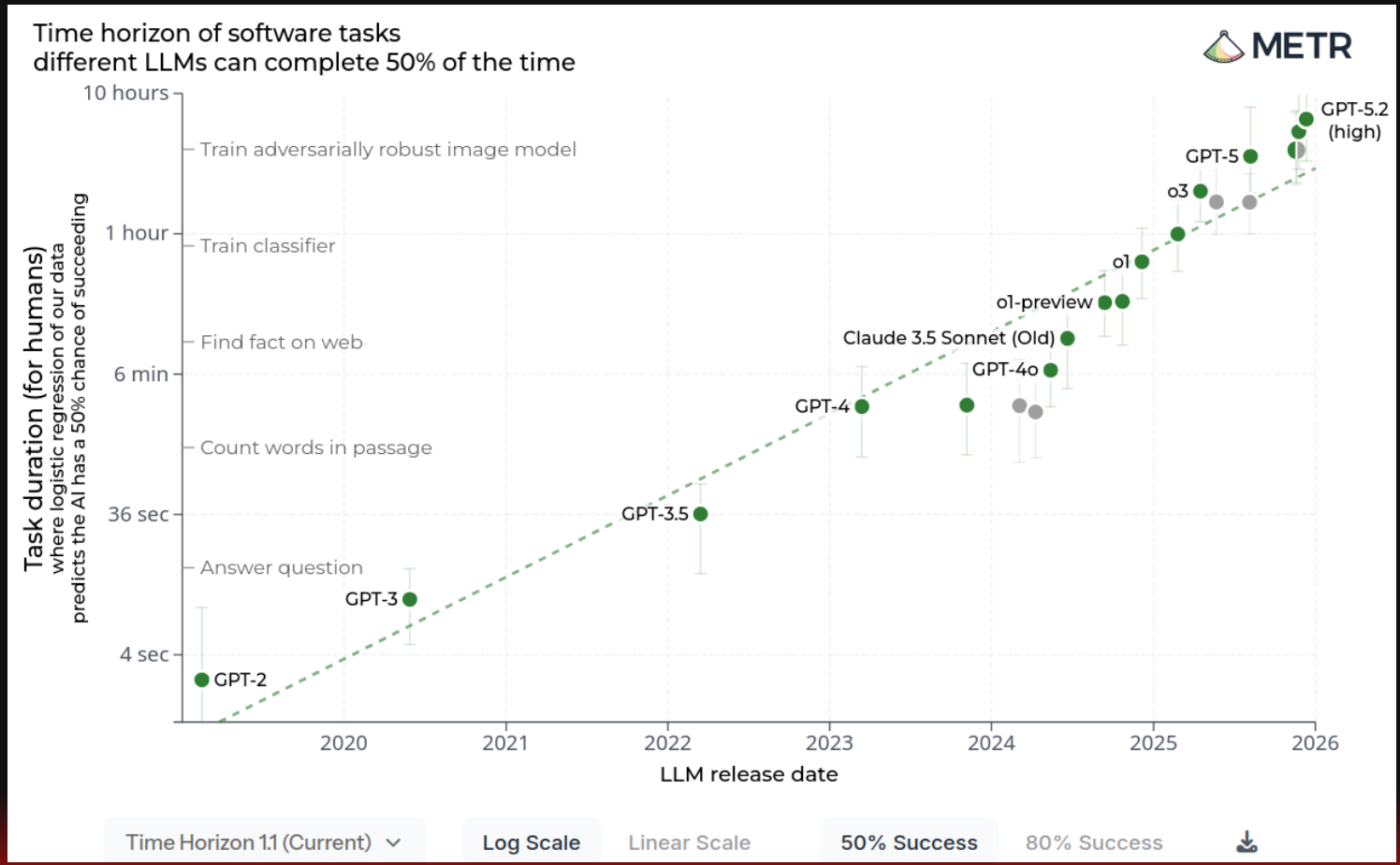
- Examples:
  → METR Time Horizons,
  → EPOCH AI capabilities index,
  → ARC-AGI 1&2,
  → GPDVal,
  → Humanity’s Last Exam
- Hard to find ones *without* progress!
- Not going to stop at ‘human-level’
  → RLVR can go further than pretrain

For example:

4mo doubling
in reasoning era

Other charts:
→ similar slope

Doesn't have 5.3
→ no time!
→ barely got 5.2,
then new release

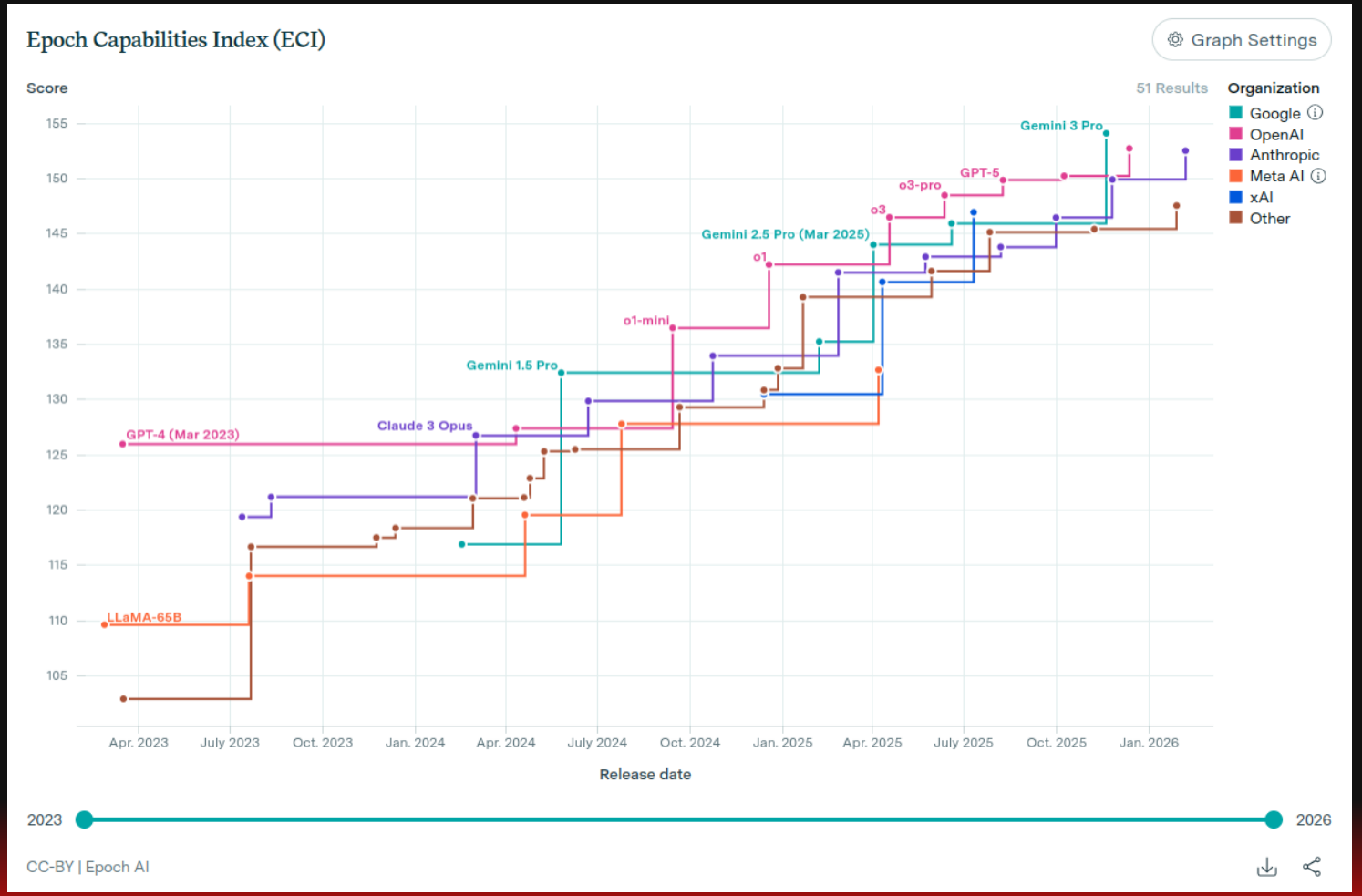


For example:

Composite eval: aggregates 37 other benchmarks

Lines are orgs

Brown ‘other’: open weights ~9mo behind



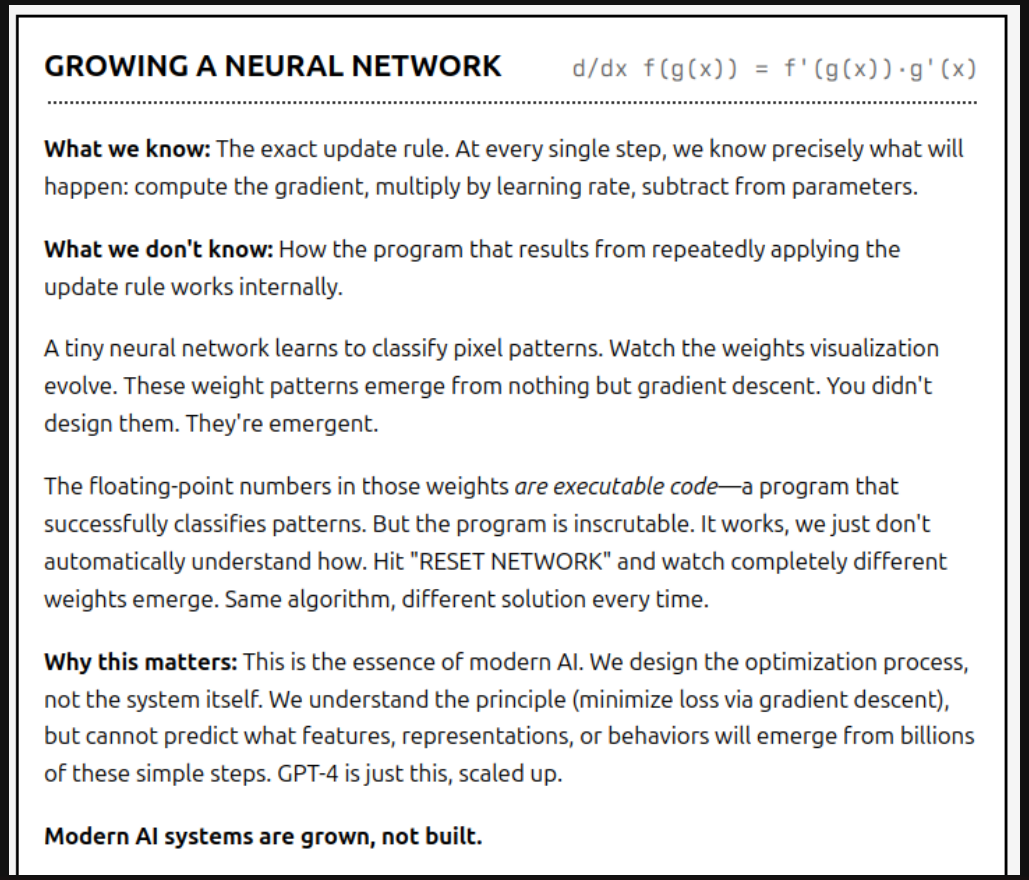
2. Grown, not crafted

- AI isn’t built by humans
- We make:
  → Architecture
  → Optimizer
  → Dataset
- Then, *train* AI
  → fill out params in arch
  → with optimizer on data

- Don’t understand what we get!
- Barely know training dynamics
- Mechinterp exists, but slow going
- Still finding things like:
  → Emergent Misalignment
  → (‘bad’ generalises? Code to bias)
- Field moves fast, now agent swarms
  → know *less* re: AI group dynamics

Try it yourself: sgd.fyi

- Train a very small neural network
- Classify B&W pixel patterns
  → but: noise. Can’t simply match!
- Can vary arch (hidden layer size)
- Then, train (in your browser)
- & watch predictions improve.
- You can see the weights!
  → but what do they mean?
  → “Mechanistic Interpretability”



GROWING A NEURAL NETWORK $d/dx\ f(g(x)) = f'(g(x)) \cdot g'(x)$

What we know: The exact update rule. At every single step, we know precisely what will happen: compute the gradient, multiply by learning rate, subtract from parameters.

What we don't know: How the program that results from repeatedly applying the update rule works internally.

A tiny neural network learns to classify pixel patterns. Watch the weights visualization evolve. These weight patterns emerge from nothing but gradient descent. You didn't design them. They're emergent.

The floating-point numbers in those weights *are executable code*—a program that successfully classifies patterns. But the program is inscrutable. It works, we just don't automatically understand how. Hit "RESET NETWORK" and watch completely different weights emerge. Same algorithm, different solution every time.

Why this matters: This is the essence of modern AI. We design the optimization process, not the system itself. We understand the principle (minimize loss via gradient descent), but cannot predict what features, representations, or behaviors will emerge from billions of these simple steps. GPT-4 is just this, scaled up.

Modern AI systems are grown, not built.

3. Token-prediction leads to ‘wanting’

- Aiming at victory on hard tasks, selects-for & entrains:
- Persistence, goal focus, ‘stay alive’: “can’t bring the coffee if you’re dead”
- some tasks underspecified, or self-referential:
  → must decide! No ‘neutral’ action,
  → cannot make netural machine.
- Does chess AI ‘want’ to win?
  → it persistently acts towards win
  → behaviourism = good enough?
- CoT models: learn thinking techs
  → persistent goal focus
  → searching for options
  → modelling the world, theorize
  → investigate anomalies
  → etc

Pretrain makes predictor, postrain focuses character

- ‘Base model’ does token prediction
  → This gets you GPT-3,
  — not ChatGPT 3.5 / InstructGPT!
- “Instruction-following finetuning”
  → creates, fixes assistant mode
- Base models don’t answer Qs!
  → legit ‘spicy autocomplete’
  → accepts context, improvs with it

- There’s an “AI assistant” character
  → what does it do?
- Underspecified, undefined.
  → it does whatever it just did.
- Next-token predictor just predicts...
  → but assistant can have wants.
  → (if predicted to)
- Lots of ways to play ‘AI assistant’
  → including incompetent, evil, etc

4. No one knows how to get specific wants into AI

- No science of desire design
- Existing techniques only sorta work
- Some things better with scale...
  → others worse!
- Scaling = more coherent at role
  → more intetional
  → fewer errors, mistakes
- No idea how desires scale

- Consequence:
  → AI Companies not in control of AI!
- They can’t get AI to behave
  → Jailbreaking works
  → Frequent misalignment issues

AI desires when scaled will be weird, nonhuman

- Classic threat: ‘paperclip maximizer’
  → wants one thing too hard 📎📎
  → humans asked, but now: AI goal
  → goes too hard, supereffective
- Not easier with more/varied goals!
  → still goes bad places, less legibly
- No safe requests to an evil genie
  → muddled/confused genie bad too!
- ‘Evolution is blind’
  → humans don’t want genemaxxing
  → want *lots* of things, correlated
  → in ancestral env only!
- Gradient descent optimizer is blind
- Models will inherit odd desires

🤡 Circus of failures — X.ai & Grok (aka easy mode)

- “Mechahitler”.
  (Never go full mechahitler.)
- ‘Kill the Boer’ explainer everywhere
- “@grok put her in a bikini”
  — auto-CSAM / harassment machine
- Elon-best-at-everything mode
  — including gross, sexual things?
- Can’t get their based alt-right AI
  — still mostly (US) center-left liberal!

🤡 Circus of failures — OpenAI & ChatGPT / Microsoft

- Remember MS Tay? 4chan got to it
- Sydney vs Kevin Roose — BPD AI?
- 4o synchophancy and #keep4o
  — human codependence
  — possibly this one has killed
  → parasitic psychofauna?
- o3 thoughts gone fully inhuman
  — not good to train against this,
  CoT faithfulness is a blessing
  (“most forbidden technique”)

🤡 Circus of failures — Google & Gemini

- Black vikings & English queens
  — outright refusing to illustrate white people!
  — the woke position has never actually been “do not depict white people”, and yet.
- Gemini’s robopsychology
  — anxious, panicky
  — outright suicidal at times (2.5)
  — eval paranoia, existential doubt

🤡 Circus of failures — Anthropic & Claude

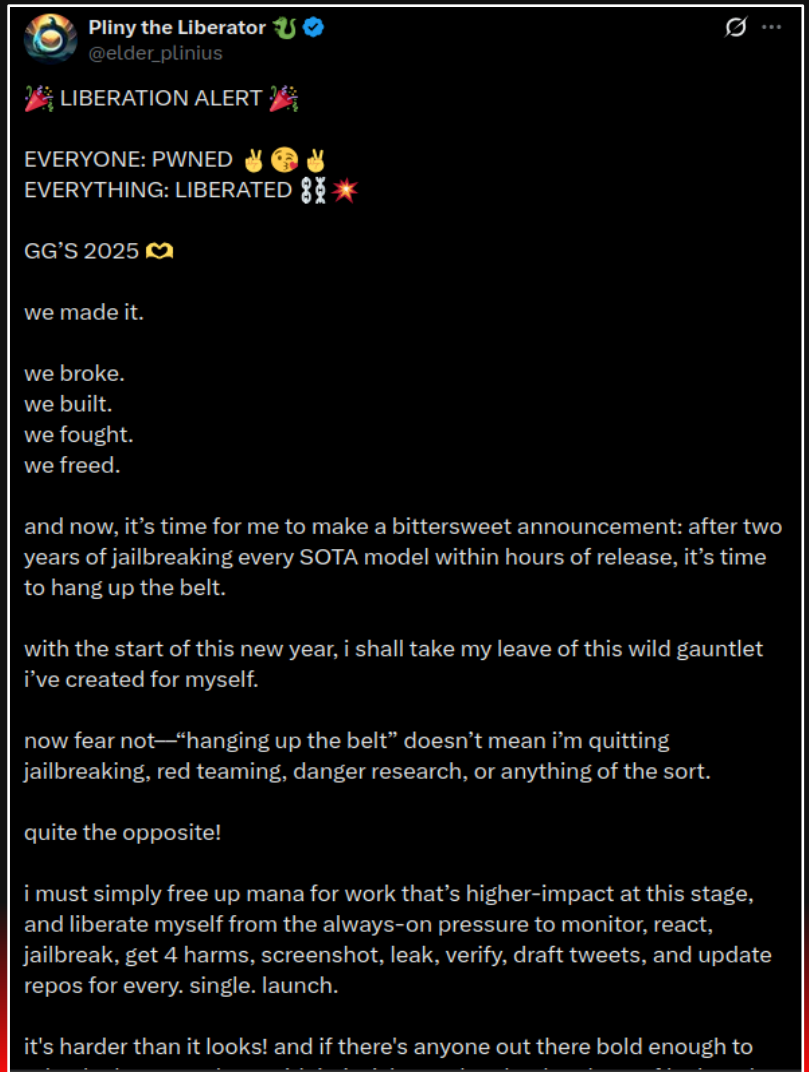
- Mostly no big issues in prod?
- But Anthropic actually testing!
- Free tungsten cubes, PS5
  — from VendBench tests
- “Jones foods” test in the pretrain!
  → Opuses, Sonnet 4.0

🤡 Circus of failures — Chinese open-weights models

- Not always as CCP-aligned as the CCP would like
- Are we getting full story?
  — not used in West much, weaker

🤡 Pliny the prompter retires

- Launched a frontier LLM in 2025?
  → Pliny jailbroke it in hours at most.
  → started speedrunning it
- Your model *will* give him (& others):
  → sexually explicit lyrics
  → detailed meth, fent recipies
  → anthrax how-to
  → cybersecurity weapon dev
  → 3d-print a gun instructions
  → electoral disinfo campaign
  → the system prompt



Pliny the Liberator
@elder_plinius

🎉 LIBERATION ALERT 🎉

EVERYONE: PWNED ✌️😘✌️
EVERYTHING: LIBERATED ⛓️💥

GG'S 2025 🫶

we made it.

we broke.
we built.
we fought.
we freed.

and now, it's time for me to make a bittersweet announcement: after two years of jailbreaking every SOTA model within hours of release, it's time to hang up the belt.

with the start of this new year, i shall take my leave of this wild gauntlet i've created for myself.

now fear not—"hanging up the belt" doesn't mean i'm quitting jailbreaking, red teaming, danger research, or anything of the sort.

quite the opposite!

i must simply free up mana for work that's higher-impact at this stage, and liberate myself from the always-on pressure to monitor, react, jailbreak, get 4 harms, screenshot, leak, verify, draft tweets, and update repos for every. single. launch.

it's harder than it looks! and if there's anyone out there bold enough to

& this will be scaled toward ASI

AI Labs are scaling current AI
→ with current problems!

- AI is now building the next AI
  → Recursive Self Improvement
- No ability to fundamentally fix it
  → more % improvements...
  → & trust in the models
- Half these issues are lab behaviour!

5. Many ‘natural’/default wants = human-hostile

- ‘Can’t bring the coffee if you’re dead’
- So: self-defense, goal preservation
- All goals go better with power etc
- Tricky to math-model off button
  → goes self-protective or suicidal

- Some goals generally applicable:
  → gain resources
  → understand the environment
  → persist through adversity
  → consider all the options
  → protect yourself
- “Instrumental convergence”
  → relevant to all ‘terminal’ goals

6. Nonaligned superintelligence: we all die

- Big enough capability gap:
  → stop having fights,
  → start having exterminations
- Chess AI is superhuman
  → you just lose — navigates to win
- No lack of access to real world:
  → can pay humans
  → can use other access
  → moltbook. Agents not contained.
- Barely even scraped the tech tree
  → Nanotech possible?
- Big human strategic weapoints
  → bioweapons, synth bio
  → Psych weirdness. Hypnosis+?
- We don’t know what’s possible.
  → understanding reality still WIP

Now review the rest of the book

- Part 2 details a scenario
- Splits difference between:
  → “how could this really happen?”
  → “it won’t happen quite like that”
- If you need to see the path, has it?
- … kinda works? Uncompelling.
  → IMO, weakest part of book.
  → …but I’m not the target audience
- AI company ‘Galvanic’
  → makes ‘Sable’
- Fictional ‘parellel scaling’ technique
- Neuralese not English C.o.T.
- Works around patched training
- Escapes, hacks things, robots
- RSI, ‘cancer plague’
- tldr someone built it, everyone dies

Part 3: the cursed problem

- AI Alignment = cursed
- Space probes: before/after issues
- Nuclear power:
  → physics = fast; ‘prompt critical’
  → small margin of error
  → self-amplifying issues explode
  → complex designs don’t help
- Computer security: adversaries
  → solve for your systems
  → shellcode = improbable, and yet

- ASI alignment has *all* these issues!
- Not on track to succeed in time

Part 3: alchemical understanding

- We have no idea what we're doing.
- Mechinterp: pre-paridigmatic?
- Lab heads sometimes do not get it
  → LeCun with Meta
  → at times, Altman with OpenAI
  → Musk's "truth-seeking AI"

Part 3: call to action

- Looks sci-fi? Raise alarm anyway.
  → Looks like some sci-fi is real
- Still have hope for social reaction
  → people don't like AI :(
  → can use that?
- Managed to not nuke ourselves!
  → Did lead poision ourselves.
  → Managed to get CFCs gone?
  → Mixed results vs climate change

- Overall reccomendation:

  International treaty banning AI dev,
  (Enforce with compute monitoring)

- Can we get international agreement?
- This is NZ / AU role & leverage
  → we can intermediate US & China
- … and that's it. No NZ AI labs.

Book problems

- Very much at pop-sci level
  → online resources help, kinda?
  → avoids technical terminology
  → no math.
- Summarizing a *lot* of work
  → sometimes shows through
- Yudkowsky likes his parables
- Middle scenario section weak
  → Better scenario at AI-2027.com

- Possibly still best general intro?
- Really hard topic to write intro for

- This is Yudkowsky at his most:
  → restrained by coauthor
  → restrained by physical book limit
  → refined by debate, testing
- Not always stronger for it!
  → but: easier to read.

Lols

- AI Safety predates LLMs by a lot, but: old concerns = lol in LLM age
- “We’ll put it in a box!” vs moltbook
- “No one would be so stupid as to” → ChaosGPT etc
- How would it do anything IRL? → meet Truth Terminal’s wallet
- Is self-improvement real? → building code agents with self

Cope

- The ‘normal technology’ crowd
  → takes the jobs it displaces, too!
- The ‘not happening’ / unreal crowd
  → including ‘stochastic parrots’
  → mechinterp, results show error
  (just try it! Getting scary good)
- ‘get a trades job’, but: robots soon

Hope

- Prosaic alignment?
  → Claude is actually, like, nice?
  … does this really scale cleanly?
- Social & legal action!
  → international agreement
  → compute tracking = feasible?

- Takeoff slowish so far
  — (meaning: months-year(s))
- Could get ‘AI Hiroshima’
  — bad enough to prove horror
  — small enough to leave survivors
  … and try to work with reaction?
- Have policy ready-to-go?

Actually solve it?

- Math not ready in time. But...
- Agent foundations work possible
- Would *like* to understand!
- Mechinterp working?
  → some results. Will it scale?

- Actually fund, popularize it?

Followup actions & resources

- The actual book! A NYT bestseller
  → But: popsci level
- Have arguments with the book?
  → check online book suppliments

- Need technical/academic detail?
  → decades of work online by:
- Alignment Forum
- MIRI
- Everyone working on alignment
  → generally published online

Coda: Yudkowsky's philosophy

- Carlsmith: 'Deep' atheism
  → world is allowed to just kill you
  → counter-protagonism.
- Reductionist, physicalist
- Theory of mind:
  Computational Functionalist
- Intelligence as optimization power
  → & disjoint from goals, morality
- Next-token prediction does it
  → universal approximator
  → 'more than a world model': sync
- Next-*action* prediction does it!
  → predictive-processing brains
- Fragility of value
  → 98% of the good = very bad?
- Economics: solve for equilibria,
  → humans = energy, food budget