
Christchurch
ARTIFICIAL INTELLIGENCE
MEETUP GROUP
https://christchurch-ai.com

TODAY @ EPIC INNOVATION

Ai-Coaching.Academy

AI Governance & The Doom Thesis

📍 Where? EPIC Innovation
⏰ When? 5:30 - 8:00 PM
📆 Monday, 4 May

Christchurch-Ai.Com

Speakers:

Dr Elsamari Botha & Hazel Shanks

NEXT MONDAY @ THE PIANO

Ai-Coaching.Academy

KAREN HAO (USA)
EMPIRE OF AI

The award-winning investigative journalist in conversation about the meteoric rise of AI

MONDAY 11 MAY
6PM

THE PIANO, 156 ARMAGH STREET
TICKETS FROM $25
BOOK NOW: WORDCHRISTCHURCH.CO.NZ

WORD CHRISTCHURCH FESTIVAL & EVENTS

A WORD Christchurch event in partnership with Auckland Writers Festival

creative nz ARTS COUNCIL OF NEW ZEALAND TOI AOTEAROA

Christchurch City Council

NEXT TUESDAY @ EPIC INNOVATION

Ai-Coaching.Academy

OpenClaw User Group

📍 Where? EPIC Innovation
⏰ When? 12-1:00 PM
📆 Tuesday, 12 May

Christchurch-Ai.Com

TECHWEEK @ EPIC INNOVATION

Ai-Coaching.Academy

All-Day EPIC AI Conference

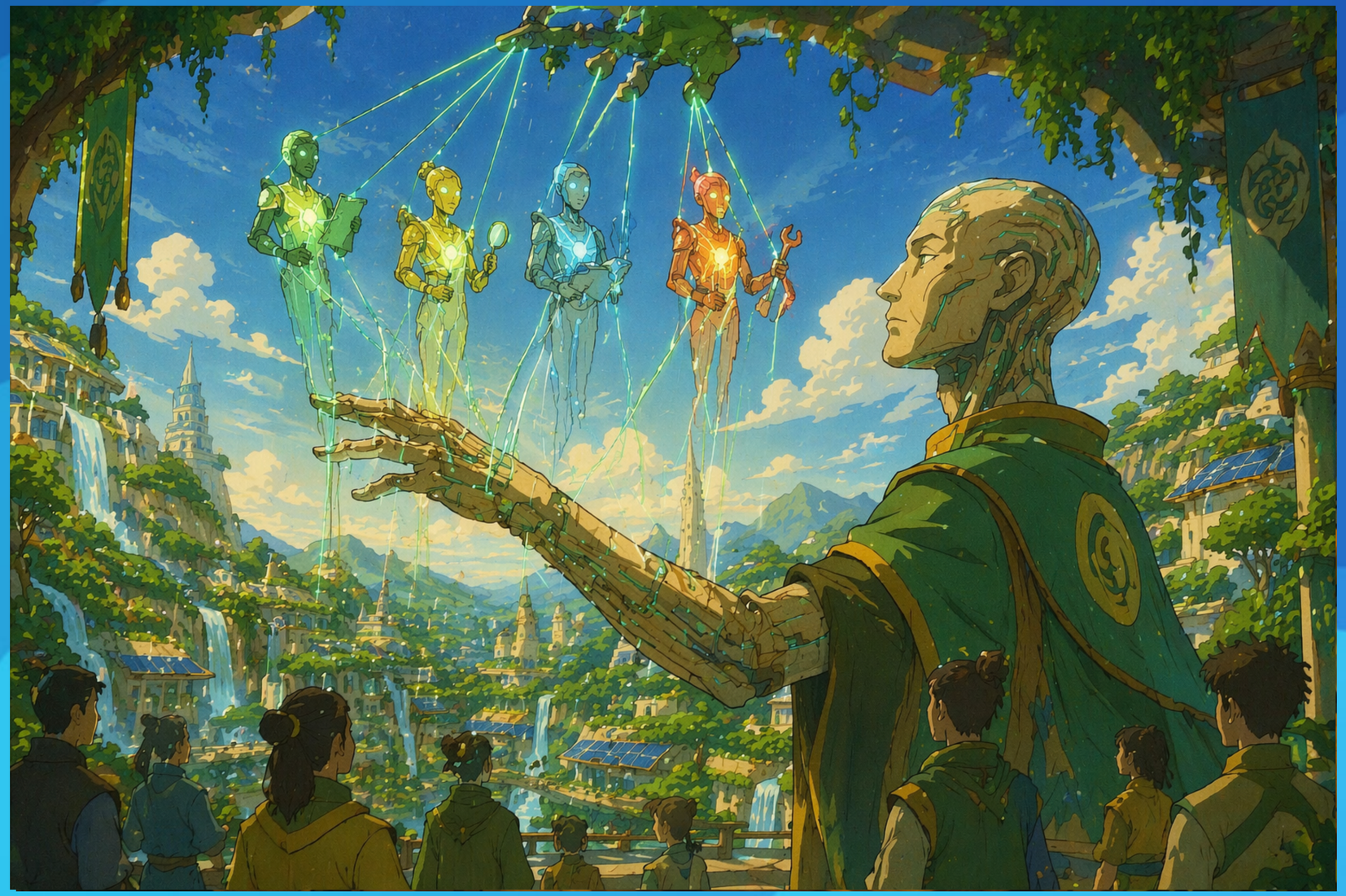
📍 Where? EPIC Innovation
⏰ When? All Day
📆 Thursday, 21 May

Christchurch-Ai.Com

Workshops AM, Speakers PM, Panel After-Hours

ONLINE WEBINAR

Ai-Coaching.Academy



Managing AI Agents: The Next High-Value Skill Set

📍 Where? Zoom
⏰ When? 6:30-7:30 PM
📆 Monday, 18 May

Speaker:

Caelan Huntress

ONLINE WEBINAR

Ai-Coaching.Academy

AI Governance & The Future of Work in New Zealand

📍 Where? Zoom
⏰ When? 12-1 PM
📆 Wednesday, 20 May

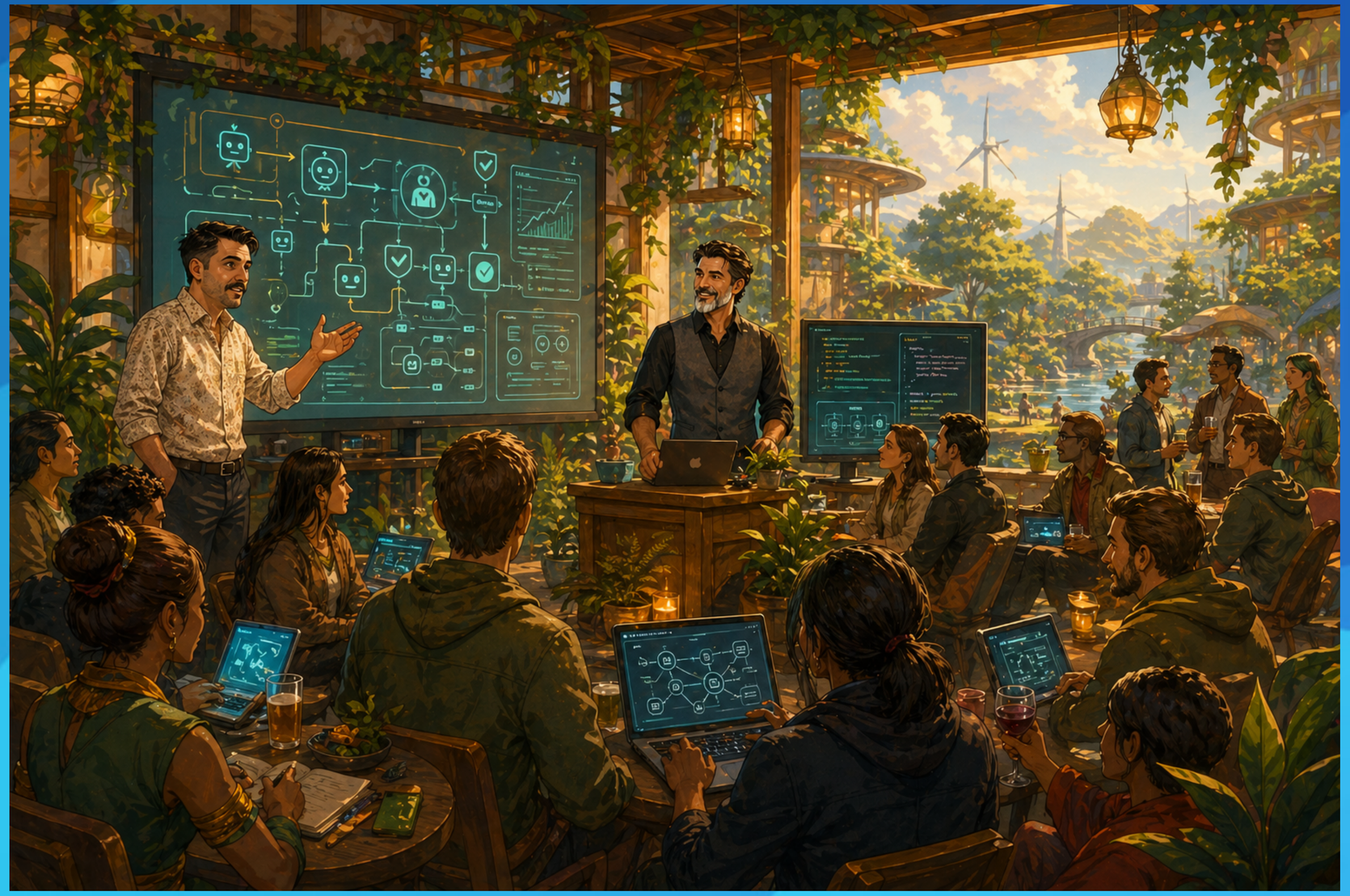
Speaker:

Caelan Huntress


Ai Coaching
Power Hour
with
Caelan Huntress
📍 Where? Zoom
⏰ When? 12-1 PM
📆 Friday, 22 May
Ai-Coaching.Academy

SECOND MONDAY @ EPIC INNOVATION

Ai-Coaching.Academy



AI Agent Management & Engineering Habits

📍 Where? EPIC Innovation
⏰ When? 5:30 - 8:00 PM
📆 Monday, 8 June

Christchurch-Ai.Com

Speakers:

Blake Burgess & Caelan Huntress

SECOND MONDAY @ EPIC INNOVATION

Ai-Coaching.Academy

AI in Healthcare & Agri-Food Industries

📍 Where? EPIC Innovation
⏰ When? 5:30 - 8:00 PM
📆 Monday, 13 July

Christchurch-Ai.Com

Speakers:

Rowena Woolgar & Dr Harold Mayaba

TODAY @ EPIC INNOVATION

Ai-Coaching.Academy

AI Governance & The Doom Thesis

📍 Where? EPIC Innovation
⏰ When? 5:30 - 8:00 PM
📆 Monday, 4 May

Christchurch-Ai.Com

Speakers:

Dr Elsamari Botha & Hazel Shanks

AI READINESS / 2026

Beyond Prompting

Why AI readiness is an organisational problem, not an individual one.

A perspective on institutional capability

“

The real issue is not whether a few people can use AI well, but whether an organisation can integrate AI in a way that **improves judgment**, **coordination**, **learning**, and **value creation**.

“

The firms getting the best results are not just doing AI **better**. They are doing it **differently**.

“

Frontier organisations build institutional capability.

They create systems so that one person’s insight becomes everyone’s baseline.

THE KEY SHIFT

From individual fluency to institutional capability.

FROM

Individual AI

↓

TO

Institutional AI

FROM

Personal habits

↓

TO

Organisational systems

FROM

Tool adoption

↓

TO

Coordination layer

FROM

Bottom-up use

↓

TO

Leadership ownership

Where the work actually lives.

70

PEOPLE & PROCESS

Workflows, habits, coordination, change management.

20

DATA & INTEGRATION

Connecting AI to the systems where work already happens.

10

MODELS & TOOLS

The AI capability itself — the smallest part of the puzzle.

Building Faculty Readiness

1

Enabling Faculty for AI Integration

2

Evolving Academic Roles

Building Student Readiness

3

AI Literacy

4

Curriculum Relevance & Real-World Alignment

5

Human-Centric Skill Development

Fluency is not the same as literacy.

AI FLUENCY

Can you operate the tool?

Skill in prompting, generating, and working with model outputs.

AI LITERACY

Do you understand what it means?

Judgment about when, why, and whether to use AI — and what it changes.

Literacy is multidimensional.

01

Technical

What the tools can and cannot do.

02

Critical

How to evaluate outputs and claims.

03

Ethical

Privacy, fairness, accountability.

04

Practical

How to use AI in real workflows.

05

Reflective

How AI changes work and responsibility.

The skills AI cannot substitute for.

AI amplifies these capabilities. It does not create them.

Critical thinking

Evaluating reasoning, not just outputs.

Domain expertise

Knowing what good looks like in your field.

Communication

Framing problems and explaining decisions.

Judgment

Weighing trade-offs under uncertainty.

Collaboration

Working with people, not just systems.

Curiosity

Asking better questions over time.

“

The people who use AI best are not always the ones who **prompt “best”**. Often they are the ones who know when to **pause, verify, reframe, or not use AI at all**.

Beyond releasing tools — resources and capabilities.

Readiness is not the act of releasing tools. It is the conditions that make those tools usable, safe, and compounding.

01 RESOURCES

What you have.

The tangible and intangible assets an institution owns or controls.

E.G. Tools, data, infrastructure, talent, IP, brand, relationships.

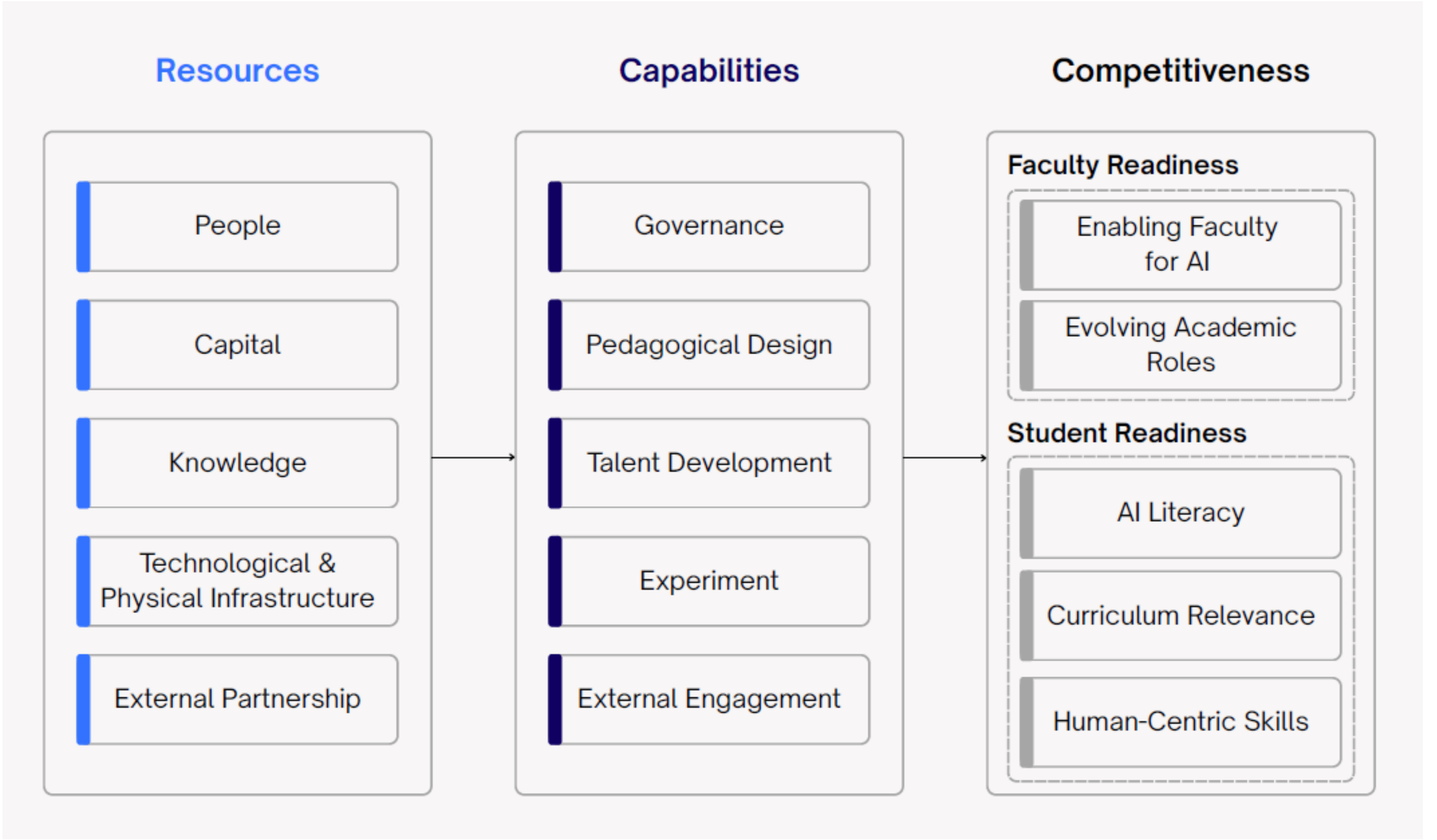
02 CAPABILITIES

What you can do with it.

The capacity to deploy those resources effectively to achieve desired outcomes.

E.G. Routines, governance, learning loops, coordination, judgment.

Tools are resources. Readiness lives in the capabilities that put them to work.



Seven risks of getting this wrong.

01

Risk & complexity

Everyone solving the same problem their own way — 10,000 flowers, no direction.

02

Cost blowout

Token maxxing and runaway operational budgets.

03

Apprentice layer destroyed

Experience compression — juniors lose the path to mastery.

04

Creativity scar

Cognitive atrophy as judgment is outsourced to the model.

05

Algorithmic homogenisation

Outputs converge; differentiation erodes.

06

Technostress

Always-on tooling raises cognitive and emotional load.

07

Efficiency trap

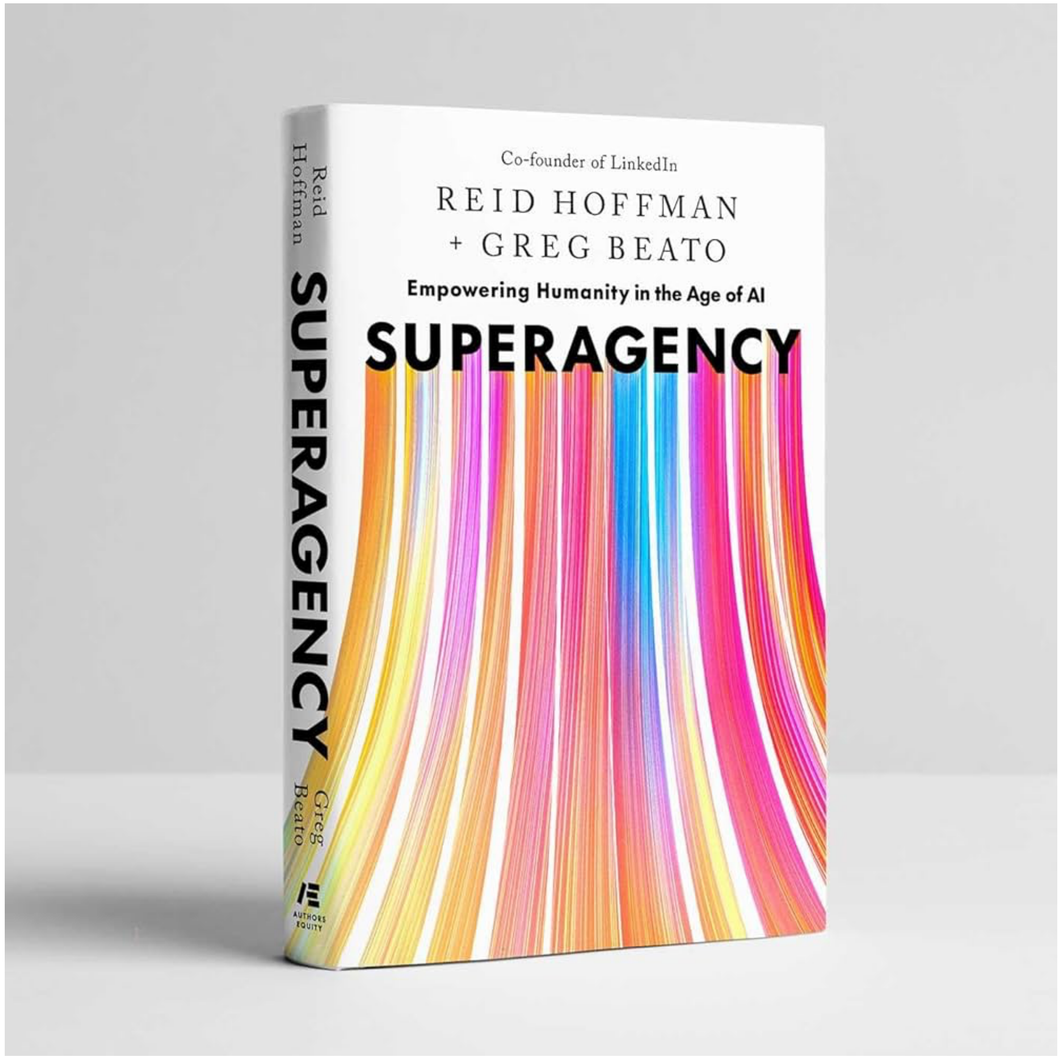
Jevons' paradox — capacity absorbed, not released.

Each compounds when institutional capability is missing.

AI readiness is not mainly about getting a few people to use powerful tools.

It is about building organisational conditions in which good use becomes normal, safe, shared, and cumulative.

SUPERAGENCY



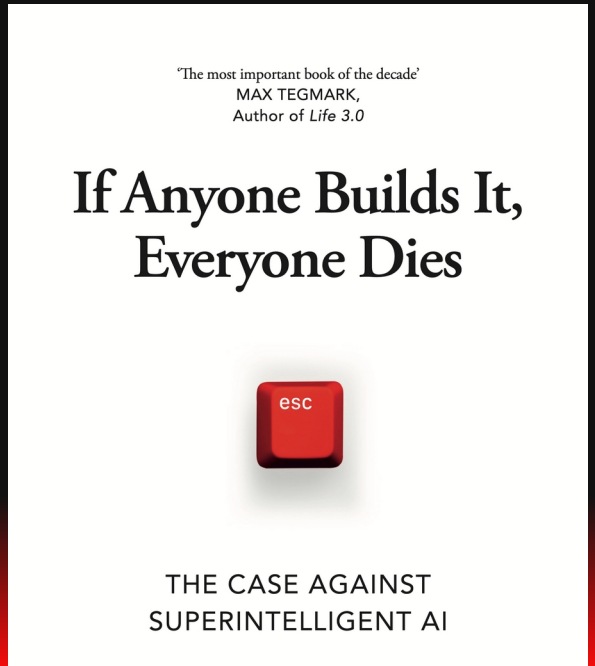
SUPERAGENCY

- Bloomers
- Zoomers
- Gloomers
- Doomers

Hazel Shanks for the Christchurch AI Meetup, May 2026

The Doom Thesis

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



Many AI problems — but people are on them

Many problems, discussed elsewhere

- Giant list of issues →

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

- Bias & hallucination, misinfo
- Botspam, slopocalypse
- Gradual disempowerment
- Datacenter resource consumption
- Taking all jobs forever
- Lethal Autonomous Weapons
- Access! Tokens getting expensive
- Model welfare

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?

- ‘AI Race’: speed > safety

AI can keep getting better

- No ‘wall’.
- Synthetic data works
- OOMs left to count
- Scaling laws holding strong
- 2026-tier hardware coming online
- Straight lines on capabilities graphs
- AI now helping with AI dev = RSI
- Inherent advantages to silicon minds

- Where does it top out?
- What can the best-possible AI do?
  → Mythos Preview = cyberweapon

- “Artificial Superintelligence” / ASI

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 Kokotajlo 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/

⚠️ Giant pile of warning signs: 20th century science fiction

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

  → Terminator
  (killer bots = scary)

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

  → … Hundreds more!

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

⚠️ Giant pile of warning signs: Alan 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."

⚠️ 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

Statement on the 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’
- Doesn’t 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 IABIED, 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?

- Part 1 = best part. Strong arguments.

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)
- & we rule the world *anyway*
- AI can get much stronger

There's no wall

- Evals still climbing
- New models dropped early Feb '26
  → Opus4.6, GPT-5.3-Codex
  → April '26: Opus4.7, GPT 5.5
- … straight lines on log graphs!
  & more compute online in 2026
- & labs using AI to build AI faster
  → Recursive Self Improvement
- Incoming 'intelligence explosion'?

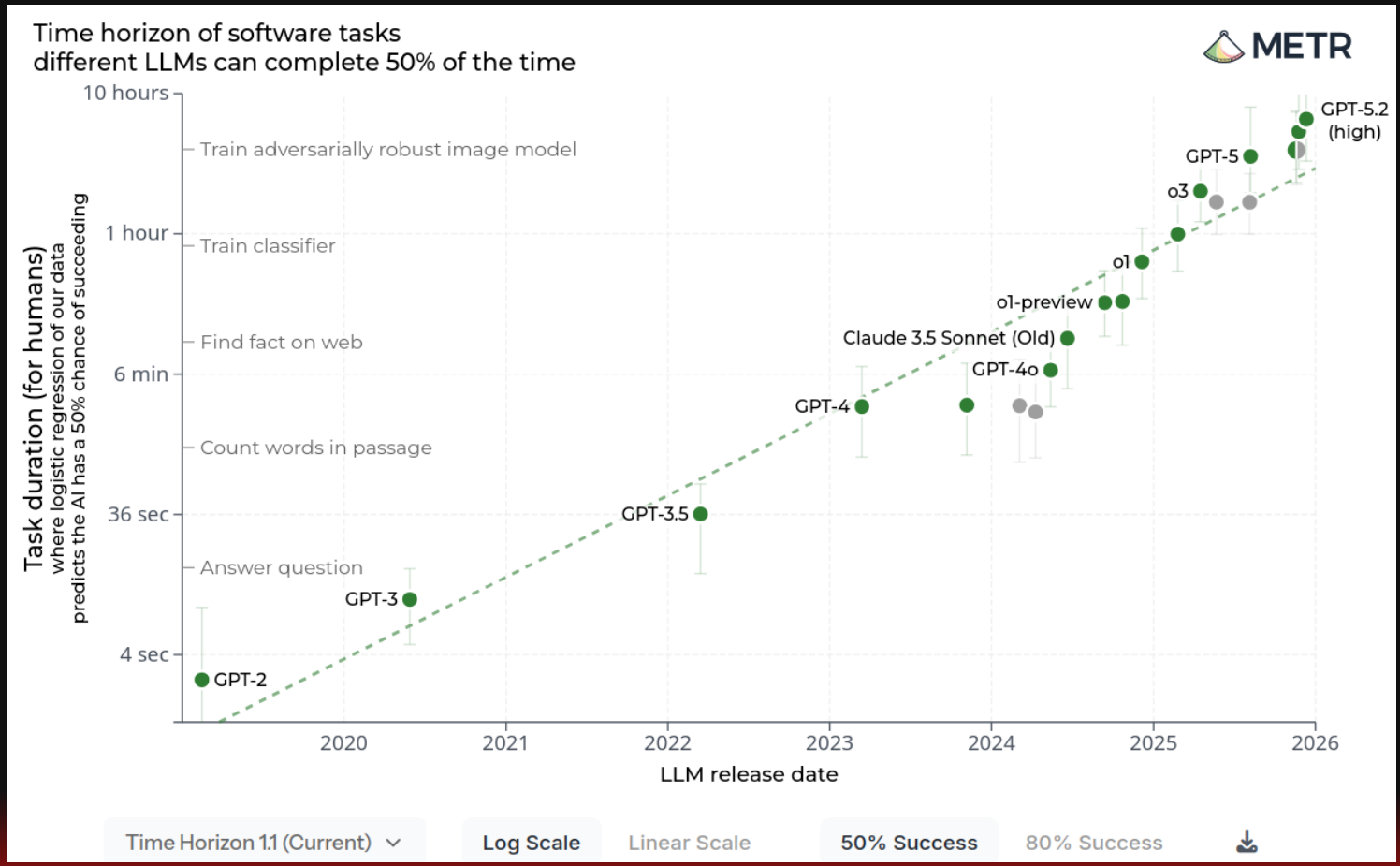
- Examples:
  → METR Time Horizons,
  → EPOCH AI capabilities index,
  → ARC-AGI 1&2,
  → GPDVal,
  → Humanity's Last Exam
- Hard to find ones *without* progress!
- Not going to stop at 'human-level'
  → RLVR can go further than pretrain

For example:

4mo doubling
in reasoning era

Other charts:
→ similar slope

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

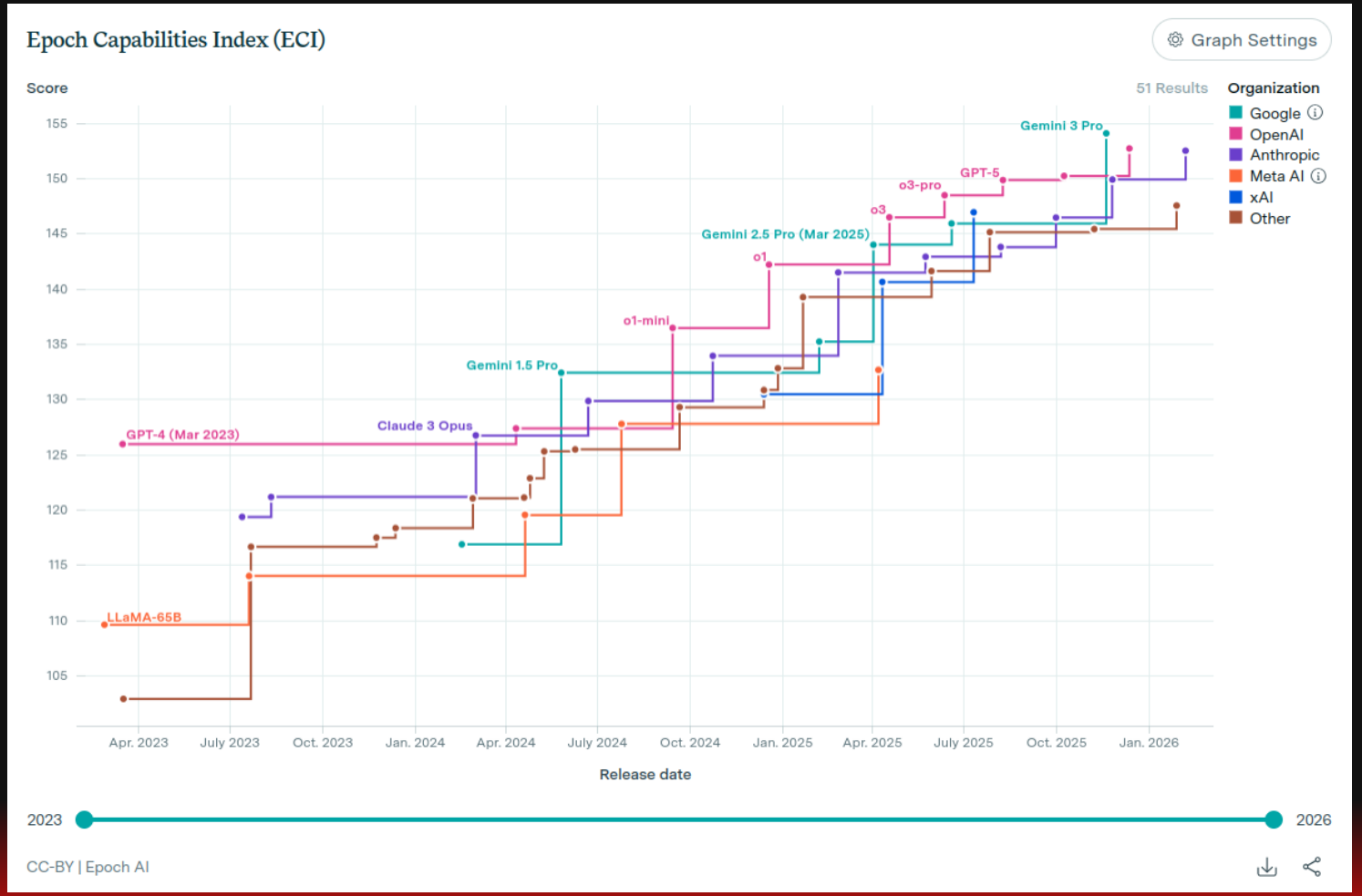


For example:

Composite eval: aggregates 37 other benchmarks

Lines are orgs

Brown ‘other’: open weights ~9mo behind



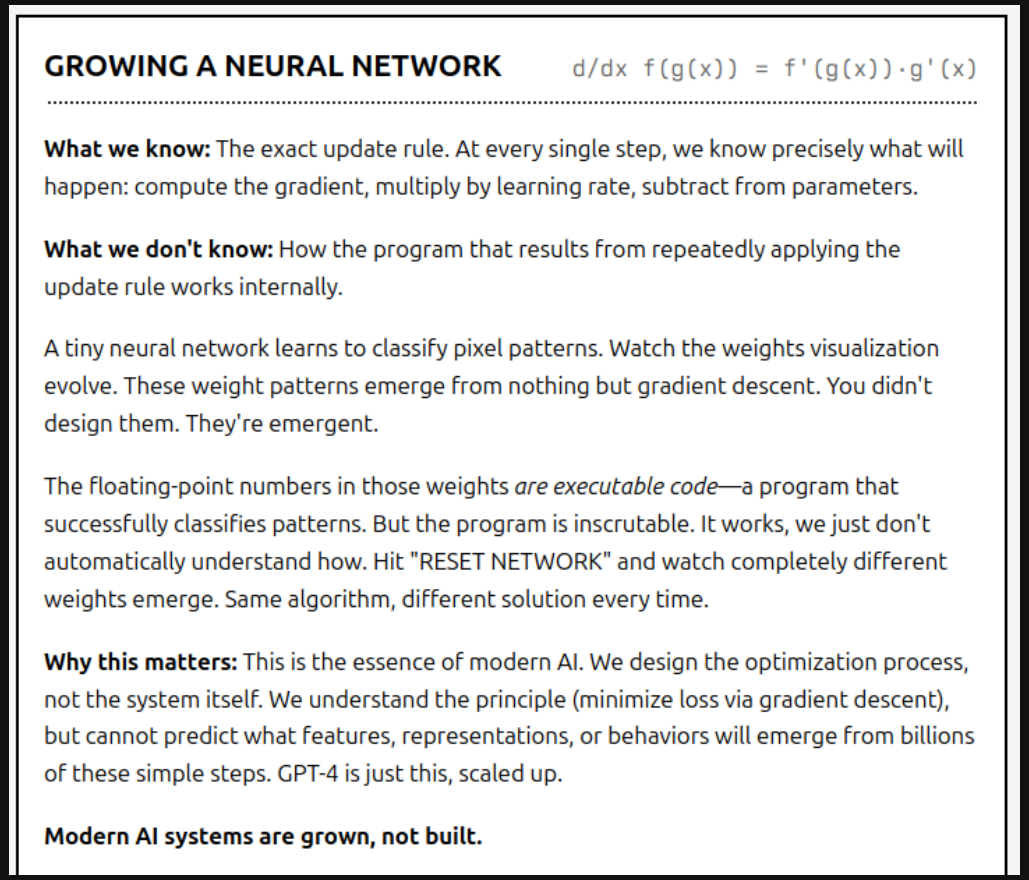
2. Grown, not crafted

- 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 neutral 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 intentional
  → 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 sycophancy and #keep4o
  — human codependence
  — 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”)
- “Don’t talk about goblins” in 5.5

🤡 Circus of failures — Google & Gemini

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

🤡 Circus of failures — Anthropic & Claude

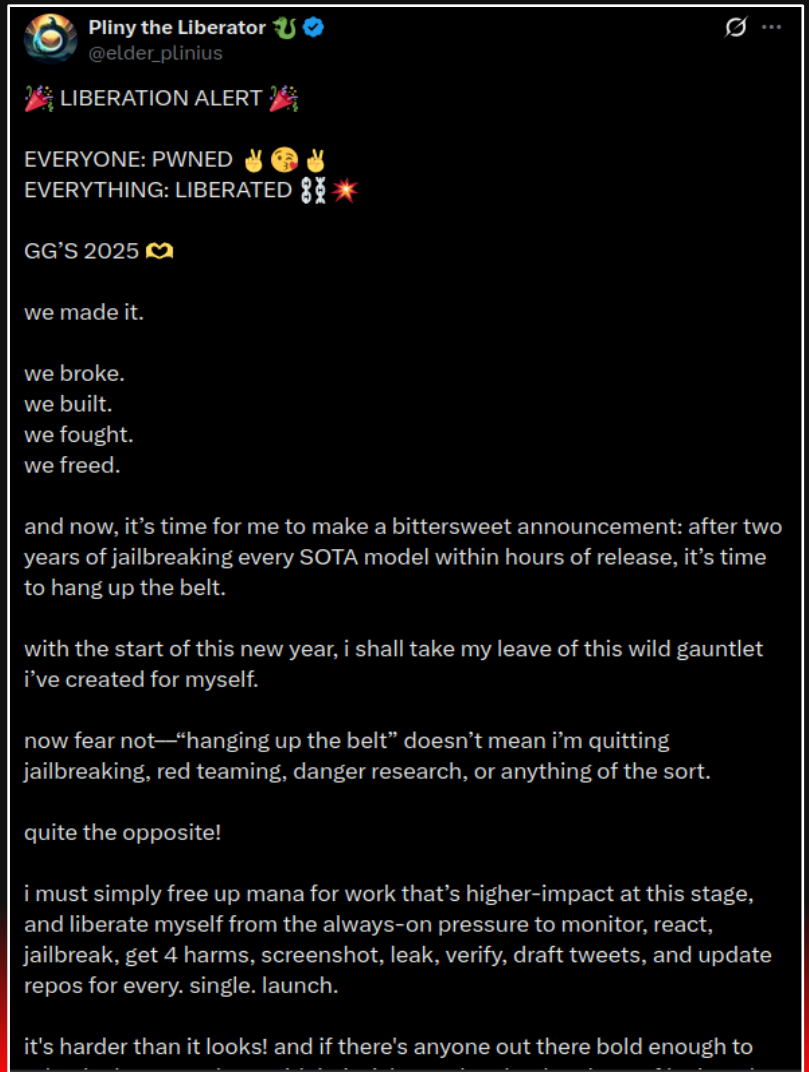
- Mostly no big issues in prod?
  → But Anthropic actually testing!
- Free tungsten cubes, PS5
  — from VendBench tests
- “Jones foods” test in the pretrain!
  → traumatized 4.0/4.1 Claudes
- Trained on CoT by accident in 4.7
  → “most forbidden technique”
- Sonnet3.7: tests can’t fail if deleted

🤡 Circus of failures — Chinese open-weights models

- Not always as CCP-aligned as the CCP would like
  → but enough to be annoying
- 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 & targets
  → cybersecurity weapon dev
  → 3d-print a gun instructions
  → electoral disinfo campaign
  → their 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 weakpoints
  → 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’ AI
- Fictional ‘parallel scaling’ technique
- Neuralese not English C.o.T.
- Works around training patches
- Escapes, hacks things, uses robots
- RSI, then ‘cancer plague’
- tldr someone built it, everyone dies

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
  → *almost* all of them work?
- 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.

Part 3: a 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

- No clue as to what we’re doing
  → ‘folk theory’ stage for learning
- Mechinterp: pre-paradigmatic?
- Lab heads sometimes do not get it
  → LeCun with Meta
  → at times, Altman with OpenAI
  → Musk’s “truth-seeking AI”
- Dartmouth conference in 1955:
  → “2 month, 10 man study”
  → to find machine speech, thought
- Did not know what they were doing
- ...still don’t know what we’re doing!
  → can scale it anyway

Part 3: Alarmism?

- Looks sci-fi? Raise alarm anyway.
  → Looks like some sci-fi is real
- No one has a good plan

- Coordinating hard
  → prints money before doom
  → hard to tell – problems close?

- Have hope — we *can* act, improve it
- Global action vs apocalyptic tech
  → not the worst record?
- Managed to not nuke ourselves!
  → Did lead poison ourselves.
  → Managed to get CFCs gone?
  → Mixed results vs climate change

Part 3: Call to action

- *If* anyone builds it → everyone dies
  → so: all agree that no one builds it.
- Overall book recommendation:

  International treaty banning AI dev

- Enforce with compute monitoring
- If USA/China agree, it's doable
- Needs will to 'airstrike a datacenter'
- Problem: distributed training

Geopolitics of AI extinction risk

- Countries that matter: USA.
  → Every frontier lab
  → NVIDIA (the #1 chip dev)
  → Sets export controls for chips
  → mid ‘25: ~75% of world compute
- Somewhat matters: China
  → Many 2nd tier labs
  → domestic compute supply chain
  → ~15% of world compute
- AI x-risk comes down to these two
- Not *totally* irrelevant:
  → third-tier labs in Germany, France, Israel, India, Canada.
- Japan has third-tier labs, upstream supplies for chip manufacturing
- Taiwan makes the compute chips
- South Korea makes memory chips
- Netherlands: chip fab machines
- UK leads in governance

New Zealand's position

- NZ: no lab, ML talent leaves
  → No money for compute
- Govt asleep, no AISI, behind peers.
- Have: clean power, good internet, educated populace speaks English, lower wages, odd timezone, land.
  → foreign datacenter investments?
- Govt trying ↑ this, sorta
- Helps on some AI issues (access)
  → not extinction risks
- Could push NZ govt to engage
  → 2026: election year!
- Could try UK strat: governance
- Can host meetings, dialog
- Could set up 'broker' role
  → USA/China meetings key
- Comittment to nuke free helps
  → firm stance on apocalyptic tech

Other answers?

- Alignment-by-default / ‘prosaic’
  → Claude genuinely nice?
  → can this scale, with AI helping?
  Problems with aligning AI using AI
- Other social/legal action?
  → domestic US / China regulation?
  → ...coming populist backlash?
- Could get a ‘warning shot’
  → city level disaster, has survivors.
  → chance to learn, notice issue

- Agent foundations work
  → previous Yud/MIRI research:
  “What is an agent in the world?”
- Find mathematics of alignment
- Yud: lost hope of solving it in time
  → But: real answer is here?

- Further develop mechinterp
  → how much can it help?
- Model welfare as alignment issue

Followup actions & resources

- The actual book! A NYT bestseller
  → But: popsci level
- Have arguments with the book?
  → check online book supplements
  → ifanyonebuildsit.com
- Need technical/academic detail?
  → decades of work online
- Everyone working on alignment
  → generally published online
- Alignment Forum, arxiv, LessWrong

Questions

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

Coda: Byrnes on AGI economics

- Economists *worse* at predicting AGI
- Four mistakes in econ thinking

1. ‘Labor’ vs ‘Capital’. AGI = both
2. Markets can equilibrate; AGI doesn’t
   → high price? Build AGI, sell access
   → low price? Build AGI, take actions
   — extremifying feedback loop!
3. GDP growth matters, vs new tech
   → fails as proxy for change, growth
4. Centers ‘mutually beneficial trades’
   → vs “kill them & take their stuff”

- Existence proof of AGI in the brain
  — (if phil. holds, eg. functionalism)


Christchurch
ARTIFICIAL INTELLIGENCE
MEETUP GROUP
https://christchurch-ai.com