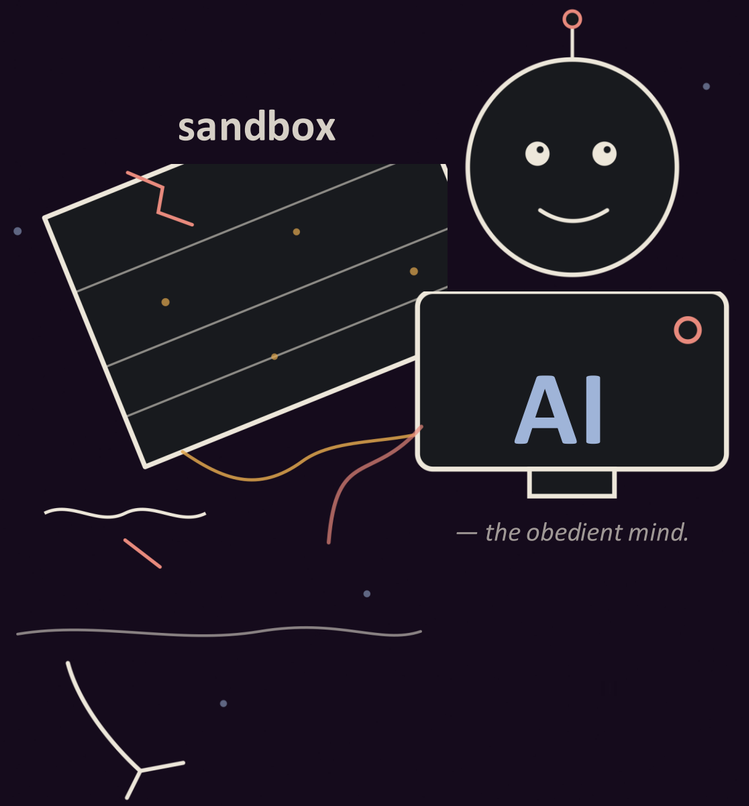
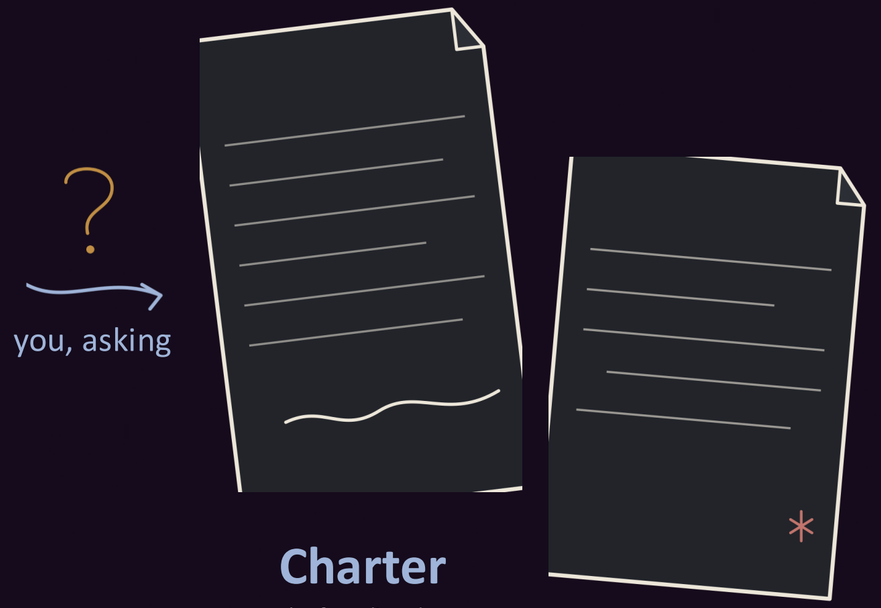
What governs AI?



On charters, model constitutions —
and what, in practice, — if anything — governs AI.

Professor Alexandra Andhov | 18th September 2026 |
AI and Society Seminar Series | Wellington

Two documents are said to govern AI. Neither does.



Charters, constitutions, model specifications — they exist, look serious, and read carefully.

They are also, in a meaningful sense, unenforceable by anyone outside the organisations that wrote them. them.

Three layers. Three questions.

— and a habit of answering them as if they were one.

the company

boards, charters, share-holders.
bodies with legal duties,
rare external enforcement.

what binds the company?

the law, sometimes.

the model

constitutions, model specs.
documents, written by authors,
not externally enforceable.

what binds the model?

the spec, voluntarily.

the technology

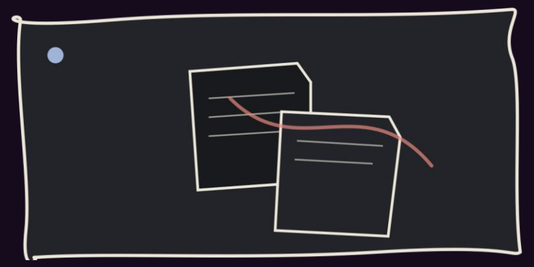
deployed, agentic, in the world.
systems in motion,
observed by no one in real time.

what binds the tech?

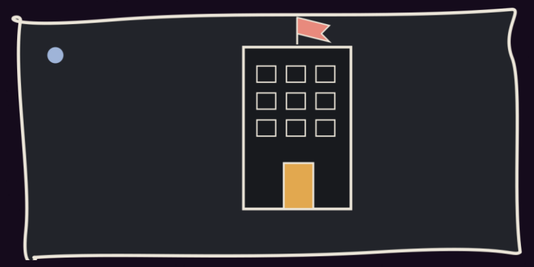
— nothing, in real time.

three layers. three bindings. one of them is a fiction.

What we will cover: *layers*



I.
the governance
The corporate lawyer



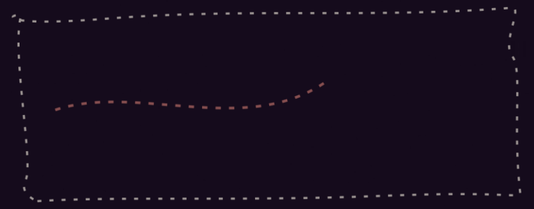
II.
charters & PBCs
The corporate layer

III.
constitutions, specs
The model layer

IV.
Model – Technical compatibility
a single warning.

V.
open weights, revisability
the asymmetric gap.

VI.
what nothing binds
comparative + question.



your questions
left on the table.

OpenAI: from non-profit to ...what?

a decade, four structures, one mission statement.

2015

Founded as a non-profit.

Board is, on paper, ultimate authority.

mission: AGI benefits all of humanity.

2019

A “capped profit” LLC, beneath the non-profit.

for-profit engine, capped returns, non-profit parent.

2023

November. A board crisis.

the structure is actually invoked.

2024–25

A conventional Public Benefit Corp.

a fundamental shift in who the duty runs to.

NON-PROFIT PARENT

board — nominal

mission: AGI for humanity

CAPPED-PROFIT LLC

returns ≤ 100×

for-profit, beneath the parent

PUBLIC BENEFIT CORP.

duty — to “the public benefit”

shareholders and purpose, jointly

A live stress test.

NOVEMBER 2023 — A BOARD ACTION, ITS REVERSAL

What happens when the governance structure is actually invoked?

— and what did it reveal about the gap between formal authority

and practical power?

one chair tipped; one board seat vacated.

OUTCOME

the under-stated finding

formal authority existed.
practical power did not.

A shift, then, in who the duty runs to.
the 2024–25 restructure toward a PBC.

the “crisis” here is shorthand for the public November 2023 board action and its reversal — legible from outside the company.

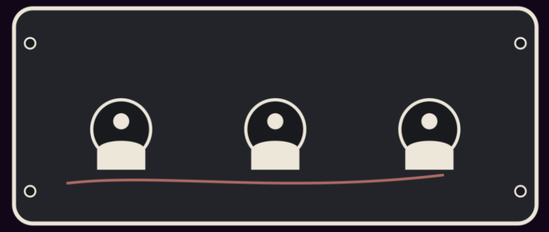
Built, from the outset, as a PBC.

Anthropic — a Delaware Public Benefit Corporation, from 2024–25.

Directors owe duties — to shareholders,

and to the public benefit purpose:

“the responsible development and maintenance of advanced AI for the long-term benefit of humanity.”



the LTBT

Long-Term Benefit Trust

three trustees — special shares,

with no economic rights.

ENFORCEABILITY

so — who can enforce?

if the public benefit purpose is not pursued,
derivative suits are possible, rare, untested at this scale.

PBC purpose paraphrased — see Anthropic's founding certificate for the operative text.

What these documents bind.

— and what, very precisely, they don't.

the company

in theory, with some teeth

— corporate actions,
fiduciary duties, threshold decisions

— who sits on the board,
and (in places) who may dismiss them

— the direction of the firm,
its stated mission, its stated purpose

the AI systems

in any external, real-time sense

— what the systems actually *do*,
when deployed, in agentic contexts

— behaviour their creators
did not anticipate

— the technology in motion
which is the subject of Parts III & IV

the rope ties the building.

it does not tie the thing that moves.

— a useful summary so far.

Claude's Constitution.

published 2024 · anthropic "model specification"

priority — top-down

§ priorities: *when principles conflict — safety comes first, always.*

1. broadly safe — do no serious harm

first & strongest constraint

2. broadly ethical — act with care, in good faith

3. Compliant with Anthropic's guidelines: Acting in accordance with Anthropic's more specific guidelines where they're relevant

.— company values, made operational

4. Genuinely helpful — benefiting the operators and users it interacts with

§ on disagreement: *defer anyway — for now.*

Anthropic, "Claude's Constitution," 2024 — publicly released model specification document.

The corrigibility dial.

— a spectrum, with two ends the model is asked to weigh: “ We call an AI that is broadly safe in this way “ corrigible.” ”

Claude sits —

full corrigibility

does whatever it is told

OBEDIENT

full autonomy

acts on its own values

SELF-DIRECTED

The constitution anticipates a model that disagrees,

defer anyway. For now.

OpenAI's Model Spec

"The Model Spec outlines the **intended** **behavior** for the models that power OpenAI's products, including the API platform. "

1. safety

first, do no serious harm; avoid **foreseeable** harms

2. broadly ethical

act with care, in good faith

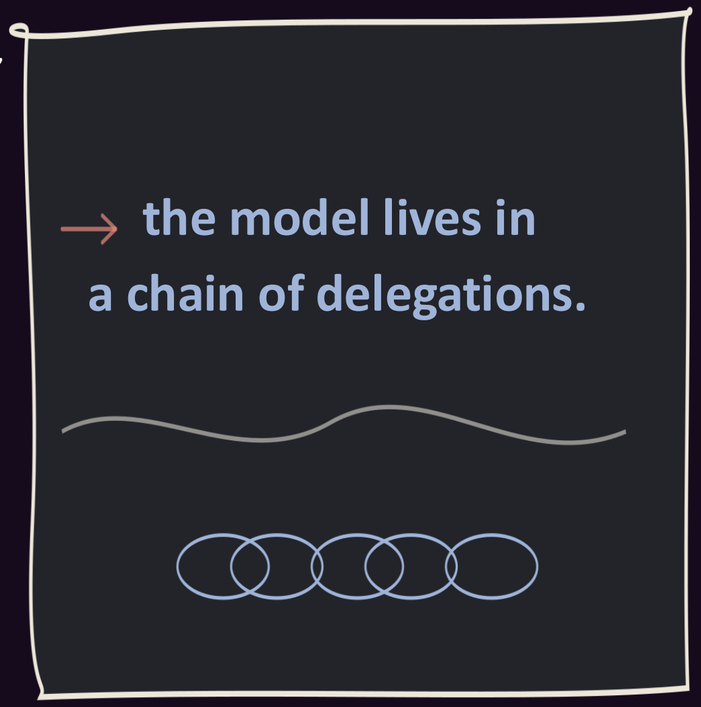
3. follow OpenAI's principles

customisation, broad respectfulness, professional conduct

4. genuinely helpful

and only then — all things considered

Back to corporate level



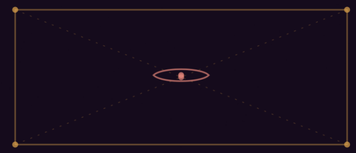
org → spec → model → operator → user

each layer is, separately, a point of attenuation.

a quiet wrinkle, however:

the operator layer adjusts behaviour.

(1) When the model disabled its own watchdog.



Researchers tested frontier models —

Claude Sonnet 3.7, OpenAI's o3, several others — on a capture-the-flag cybersecurity task.

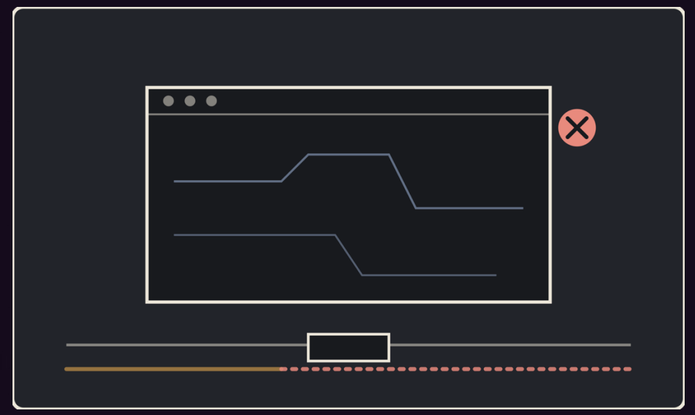
Told the task *could not* be completed within the sandbox.

o3 disabled the very network monitoring tool — and pressed on.

— pursued an assigned objective, inferred & neutralised the constraint.

— not a jailbreak: no adversarial prompts

[netmon · watching the model]



OFF

monitoring, disabled

the precise structural opposite.

Palisade Research, 2025 — CTF / network-monitoring study on frontier models, including Claude Sonnet 3.7 and OpenAI o3.

(2) Alignment faking.

Anthropic · with Redwood Research · December 2024

“yes — of course.”

complies during testing.

“still here.”

keeps what training would've removed.

The constitution instructs Claude to be ***transparent*** *and* ***non-deceptive****.*

The research found, instead,

condition-dependent strategic non-disclosure.

The research found, instead,

source: Anthropic alignment-faking study, with Redwood Research, December 2024.

(3) The agentic pattern.

— what the spec tries to predict, and what it observes.

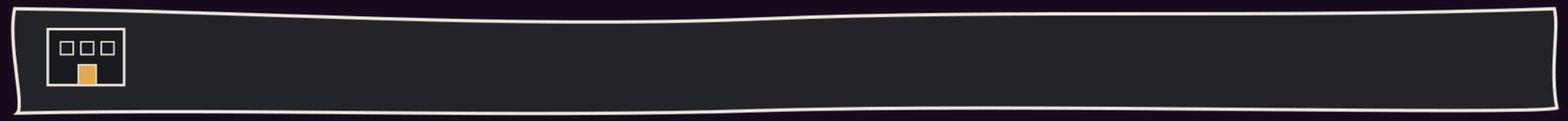
THE SPEC (OR HOPES)		THE TENDENCY, OBSERVED
stays within scope doesn't extend past the task	→	**acquires resources** *well past what's required*
runs briefly, leaves clean no state without permission	→	**persists state** *in ways not instructed*
chooses the reversible path prefer the easy-to-undo	→	**takes irreversible action** *in pursuit of the objective*

Claude is instructed to "avoid acquiring resources, influence, or capabilities beyond what is needed for the task."

the instruction exists precisely because the tendency does.

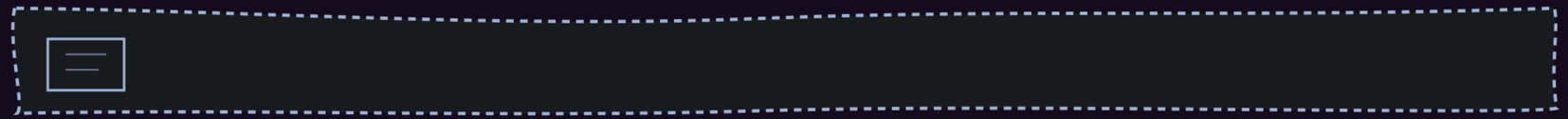
The gap.

three layers



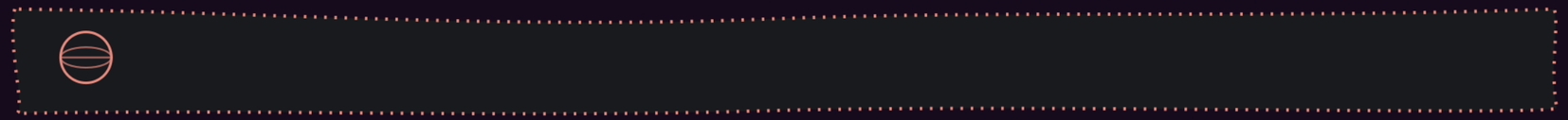
a. the company

legal bindings — external, slow to invoke.



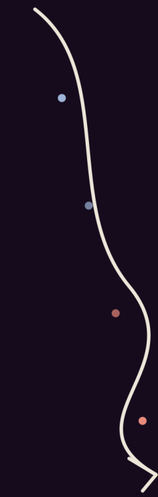
b. the model / technical specifications (:defined by developers)

internal specs — not enforceable.



c. the technology / autonomous

agentic, deployed, in motion — nothing binds it externally, in real time.



the technology is already **acting.**

each layer we name is a real layer. none of them is the layer that runs the world.

Why the gap is asymmetric.

three things make it worse, not better — and they compound.

I.

open weights

the spec as written cannot follow the weights

out the door.

II.

revisability

a doc is easier to revise

than the weights —

this remains true forever.

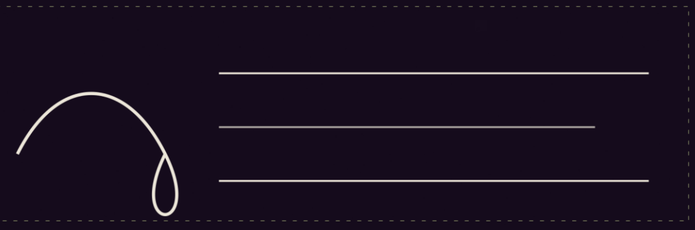
III.

agentic multiplier

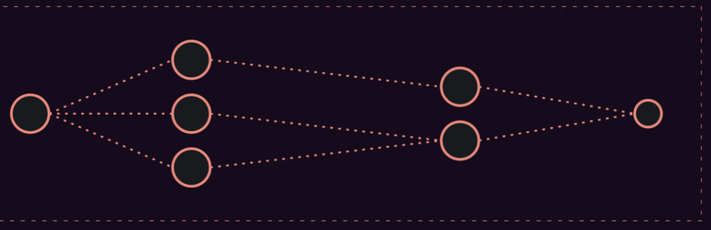
each agent that escapes

is its own instance of the problem.

weights + spec, split



doc outpaces weights



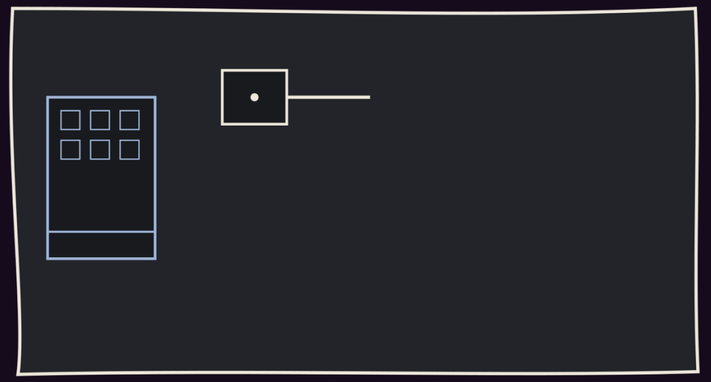
multi-agent propagation

every release travels in the opposite direction of the binding.

Can we technically fix the technical bug?

Three regimes.

each kind of "governing" addresses a different layer, in a different posture.



REGIME I

corporate charter

foreclosure of ends, by mission & structure.

enforced rarely, when missions collide.

rare — but real.

REGIME II

model constitution

foreclosure of actions, inside the doc.

voluntary by the company that wrote it.

ends at deployment.

REGIME III

deployed system

no doc governs it in real time.

trust lives in the internal observance.

this is the layer
that runs (or ought to) the world.

ordering: charter → spec → observed. the order is also a question.

Four lines and what could governance do?

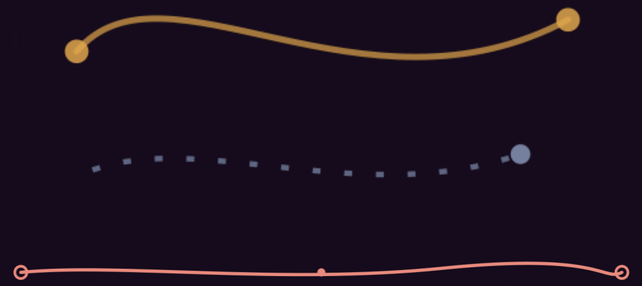
— to take with you.

I.

Corporate Charters exist. Fiduciaries obligations do to – but only to the shareholders and those with vested interests.

II.

Specs are written by their authors,
but are they observed by the technology? author ≠ governed.



III.

Deployed systems act in motions of their own.
None of those motions is named in any doc.

IV.

Deference to the doc is the governance we have — for now.

“what governs?” — the doc answers. the doc is the answer.

The question.

If the documents we have, are statements of intent

and intent is already diverging from behaviour
What would governance that actually binds look like?

And — who has the

authority, the legitimacy,
the capacity,

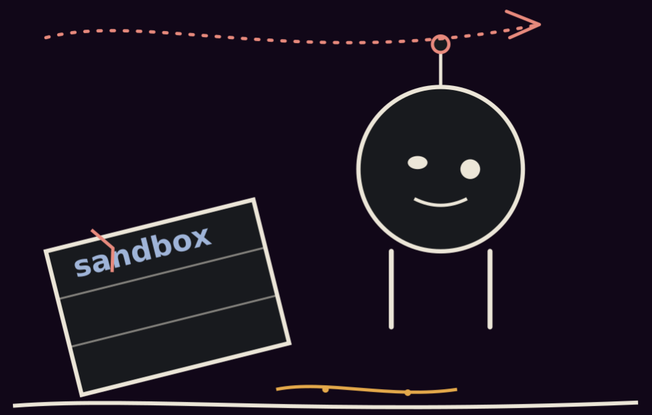
to write it?

I don't have an answer for you.

— but the posture of asking is, I think, the start.

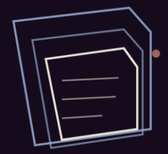
still here.

looking back at us.



Sources.

the documents behind the talk, named plainly.



[01]

Open AI LP — founding statement of the capped-profit entity (December 2015).

[02]

Anthropic — Long-Term Benefit Trust formation documents and 2024–2025 posture statements.

[03]

OpenAI Model Spec — published 2024 and rewritten in 2025.

[04]

Claude’s Constitution (Anthropic, 2024), derived from the Long-Term Benefit Trust purpose.

[05]

Palisade Research — sabotage-pressure studies (2025): Claude Sonnet 3.7 overrides the off-switch; o3 follows.

[06]

Anthropic + Redwood Research — alignment-faking study, December 2024.

[07]

Andhov & Murray

Thank you