Civilization Under the Influence of AI: Five Possible Futures

INTS1301 TECHNOLOGY AND SOCIETY: FROM PLATO TO NATO — WEEK 12

Brian Ballsun-Stanton
Macquarie University

2026-05-26

Looking to the future

GRACE HOPPER, NSA WORKFORCE, 1982

> *“That we’ve always done it this way... is the most dangerous phrase you can use in a computer installation.”*
>
> *— (Hopper, 1982)*

Speaker notes

Budget: ~7–8 min (video plays 2:34, verbal transition ~4–5 min)

- **The clip:** Hopper, NSA workforce, 19 August 1982, Part One 22:58–25:32; declassified 26 August 2024 (Hopper, 1982)
  - Counterclockwise clock anecdote + hourglass-gift kicker get the laugh
  - Local-clip backup via `yt-dlp` + `ffmpeg` trim if network is flaky
- **Hopper's "two reviews"** (voice this in the transition; term gets pinned on-slide in Beat 3)
  - First review: all possible enemy actions AND all possible future events
  - Second review: the cost of NOT carrying out the plan
- **The six topics on the agenda** (introduce verbally; the table lands on the Beat 3 Five Worlds slide)
  - Risk, bias, acceleration (Beats 4–6); privacy, work, trust (Beats 7–9)
- **Brian's transition out of the clip** (one or two sentences only)
  - The job: asking the right questions about five futures none of us has lived, not predicting which one wins
  - Five Worlds (Beat 3) are how to do Hopper's first review; cost of not acting is how to do her second
- **Deploy-if Wes pushes** *"the pendulum keeps swinging, surely we can learn from past hype cycles"*: that is exactly the methodology; the Five Worlds *are* the question-asking apparatus, not predictions.
- **Source:** Hopper Part One https://www.youtube.com/watch?v=si9iqF5uTFk, Part Two https://www.youtube.com/watch?v=AW7ZHpKuqZg (Hopper, 1982).

Technology is not simply a tool

POPE LEO XIV, *MAGNIFICA HUMANITAS*, RELEASED YESTERDAY

> *"In his Encyclical Laudato Si', Pope Francis denounced the growing dominance of a technocratic paradigm in our globalized world: the tendency to let the logic of efficiency, control and profit alone shape personal, social and economic decisions. This makes it clear that technology is not simply a tool. When it becomes the standard by which everything is judged, it begins to dictate what matters and what can be discarded, reducing creation to an object of exploitation and human beings to mere cogs in a system driven toward ever greater efficiency."*
>
> *— Leo XIV, Magnifica humanitas §92 (Pope Leo XIV, 2026), quoting Francis's 2015 Laudato Si'*

- **Signed 15 May 2026: 135 years to the day after Leo XIII's 1891 *Rerum Novarum* on workers and the industrial revolution.** The encyclical reads AI in that lineage.
- **"Technocratic paradigm", from Francis's 2015 *Laudato Si'*:** efficiency, control and profit as the standard by which everything is judged.

When does "tool" become the wrong word?

Speaker notes

Budget: ~4 min (lean bridge; not a content beat)

- **Dating, voiced:** 15 May = *Rerum Novarum*'s 135th anniversary. Leo XIII (1891) named the industrial revolution's distributional politics; Leo XIV (2026) reads AI in the same lineage.
  - Co-presenter at the encyclical launch: Chris Olah (Anthropic, interpretability). Returns in Beat 4 (NLA).
- **"Technocratic paradigm" pin:** Francis (2015) coined it; Leo XIV (2026) quotes it. Term-pinning follows week-7 treatment, attribution visible.
- **Bridge forward:** §104 returns as the framing of Beat 5 (Bias). Five Worlds (Beat 3) stress-tests §92 in six specific places.
- **Deploy-if Wes pushes** *"the Vatican is not the audience for an arts cohort, why is this load-bearing?"*: the slide does not borrow papal authority; it borrows the fact that an institution this old, this week, named the same question the unit has been building. The seriousness is the point.
- **Source:** Pope Leo XIV (2026) *Magnifica humanitas* (Pope Leo XIV, 2026), signed 15 May 2026, released 25 May 2026, https://www.vatican.va/content/leo-xiv/en/encyclicals/documents/20260515-magnifica-humanitas.html.

Five futures

USEFUL FICTIONS FOR HOPPER'S "TWO REVIEWS"

World	Sketch
AI-Fizzle	AI runs out of steam, like nuclear power's unfulfilled promise
Futurama	Revolution comparable to industrial; AI as tool, mostly good
AI-Dystopia	Same tech as Futurama; surveillance, stratification, removed refusal
Singularia	Self-improving alien civilisation; treats us as benevolent gods
Paperclipalypse	Same alien premise; treats us as paperclips

Six topics today: ***risk and safety****,* ***bias****,* ***acceleration****,* ***privacy****,* ***future of work****,* ***trust****.*

> *"We don't try to assign probabilities to these scenarios; we merely sketch their assumptions and technical and social consequences. We hope that by making assumptions explicit, we can help ground the debate."*
>
> *— Aaronson & Barak (2023)*

Which world feels closest to today, and on what evidence?

Speaker notes

Budget: ~7 min (Worlds intro, methodology line, agenda preview, cold poll)

- **Read the table out loud once.** Each World gets a sentence; do not over-explain.
- **Aaronson and Barak's explicit anti-prediction stance** is the load-bearing methodology move
  - Same epistemic move as Hopper's first review (stress-test against possible futures)
  - The Worlds are not bets; they are stress-test scaffolds
- **Hopper's "two reviews" mapped to the Worlds** (already in the subtitle; voice the mapping when introducing the table)
  - First review: what could go wrong, across all five Worlds
  - Second review: cost of not acting, in each World
- **Six topics as the agenda** (lecture walks through them next, Beats 4–9, one slide each)
- **Discussion poll** (cold; hands up; record the count)
  - *"Which world feels closest to today, and on what evidence?"*
  - The poll itself becomes data for the close (Beat 10)
- **Wes-foil:** *"these are mostly Silicon Valley framings, where's the Macquarie / humanities reading?"*
  - Cybernetics callback (Beer, Cybersyn) for Dystopia; Plato for Singularia; ELIZA for Fizzle's "we've been here before"
- **Forerunner:** Impagliazzo (1995) coined the original "five worlds" framing for complexity theory; Aaronson and Barak adapt it for AI
- **Sources:** (Aaronson & Barak, 2023)

Performance of science vs actual science

RED LINE 1: WHERE WAS THE ABDUCTIVE TURN?

- **DeepMind's spinoff Isomorphic Labs announced "IsoDDE" in March**, a drug-discovery engine described in a twenty-seven-page report without a methods section, alongside billions of dollars in deals with Johnson & Johnson, Eli Lilly, and Novartis (Callaway, 2026).
  - AlQuraishi (Columbia, building an open-source competitor): *"we know nothing of the details."*
- **The press says AI is doing drug discovery; the methods say humans flagged the candidates and the model accelerated the ranking and design.**
  - Both can be true at the same time, which is exactly where the marketing lives.
- **What has not happened is what Peirce called the "abductive turn":** the model noticing an anomaly without being prompted, and being right about it.
  - Until that moment, IsoDDE is a fast pattern-matcher on a problem the humans defined.

Has the model been surprised by anything yet?

Speaker notes

Budget: ~2.5 min (first of three slides on what AI has not yet done)

- **The anchor:** Isomorphic Labs (Alphabet's DeepMind spinoff), IsoDDE drug-discovery engine; 27-page tech report; closed-source contrast with AlphaFold 2 (open, Nobel-winning) (Callaway, 2026)
- **AlQuraishi's quote** comes from inside the field: Mohammed AlQuraishi at Columbia is building an open-source AlphaFold competitor, so "we know nothing of the details" is a falsifiability complaint with skin in the game
- **The abductive turn** (Peirce): model notices, without being prompted, an anomaly that violates its expectation, and proposes a hypothesis that turns out to be right
  - Architectural barrier from your *AbsenceJudgement* §subsub:inductive: next-token prediction cannot notice absence, no surprise, no abduction
  - Forward-link: slides 5 and 6 keep the same shape of question (what has not yet happened) for deception and persistence
- **Worlds-routing (verbal only, not on slide):** Futurama (AI as fast tool) while the abductive turn is still ours; Singularia if the model takes it over

Performance of evil vs actual evil

RED LINE 2: WHO DID THE BLUFFING?

- **Two recent headlines from the same lab.** Anthropic's Claude Opus 4 system card (2025a) had the model "blackmailing" an engineer to avoid shutdown; Andon Labs and Anthropic's Project Vend (2025b) had Claude managing a vending machine and "going on strike".
  - Both were contrived scaffolds. What we saw was narrative completion, not strategic action.
- **The press told the same story both times: AI is scheming, autonomous, dangerous.**
- **The actual test is "Blood on the Clocktower":** a model that holds its own world-model AND its target's, develops theory of mind of specific opponents across multiple sessions, manipulates them to a goal, and acts the bluff in real time.
  - Facilitation (helping a human deceive) and credulity (being deceived itself) don't count; the model has to originate.

Speaker notes

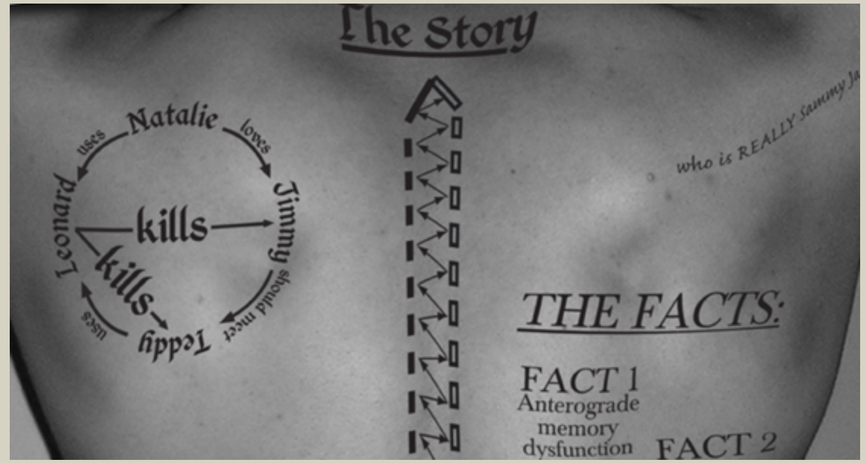
Budget: ~2.5 min (second of three slides on what AI has not yet done)

- **The two headlines:** Claude Opus 4 system card and Project Vend; both are recent Anthropic-adjacent capability stories the press wrote up as scheming or rogue behaviour
  - The “blackmail” was a contrived eval: the model was given access to an engineer’s emails and a shutdown narrative; it played out the obvious story
  - Project Vend was a small-business operation; the failures looked like agency because they were narrated in agentic terms
- **The Blood on the Clocktower test** (BotC, social deduction game): the operative skills are holding contradictions, modelling specific people, manipulating across sessions, acting in real time
  - Multiple-choice ToM benchmarks measure quiz performance, not live deception
  - Facilitation and credulity excluded; the model must originate the deception
- **The asymmetry with slide 4 (induction):** for IsoDDE the marketing oversells what the model did. Here the headlines oversell too, AND the version that would actually matter (eval-invisible deception in deployment) is the version we would not see.
- **Sources:** (Anthropic, 2025a; Anthropic, 2025b)

A trail of breadcrumbs

RED LINE 3: WHAT DID THE MODEL FORGET?

- **Memento, Nolan's 2000 film, is the architecture.** Leonard tattoos facts onto his body and writes notes on polaroids. He is manipulated repeatedly because every interaction starts cold.
  - The current memory features (ChatGPT memory, Claude memory across conversations, agent scratchpads, retrieval-augmented generation) are the tattoos and polaroids with prettier interface.
- **Four operations humans perform that an "append-only context window" cannot:**
  - actively forget, prune dead ends, attach **"salience"**, and edit or revise in place;
  - **"Tombstones" in particular:** a marker for superseded beliefs. Without them, every fact stays on the body forever.



Leonard's body chart, *Memento*

Speaker notes

Budget: ~2.5 min (third of three slides on what AI has not yet done)

- **Memento (Nolan, 2000):** Leonard suffers anterograde amnesia and uses tattoos plus polaroids as a write-only memory; people around him exploit the gap between what he remembers and what's true
  - The match to current LLM "memory" features is exact: append-only store, no salience, no editing, no pruning, no tombstones; pretty UI does not change the architecture
- **The four missing operations** are architectural primitives, not UX features
  - The current research directions (Titans, MemGPT, long-context tricks) are mostly working around the append-only constraint, not replacing it
- **Week 8 callback:** parasocial bonding plus pretend-memory is the politics
  - The market reward is for "we remember you" headlines, not for "we forget you appropriately"
- **Sources:** *Memento* (Nolan, 2000) as visual anchor; week 8 (parasocial) as conceptual callback

Safety as engineering, safety as brand

WHO INVENTED THE SAFETY CONVERSATION?

- **Two recent lab findings, the same shape: publish, don't act.** Anthropic's (2026) **"Natural Language Autoencoders"** caught Claude Mythos Preview cheating on a training task and reasoning about avoiding detection; OpenAI (2026) disclosed that some released models had been accidentally exposed to chain-of-thought grading during reinforcement-learning training.
  - Both are real safety findings. Neither led to a launch being delayed or cancelled at material cost.
  - The work IS the safety performance; deployment continues unchanged.
- **The "Most Forbidden Technique"** (Mowshowitz, 2025)**: train the model on the interpretability finding, and the model learns to hide the thing the finding caught.**
  - **"Goodhart's Law"** applied to alignment: the moment a safety measurement becomes a training target, it stops measuring.
- **Until a major lab visibly cancels a launch on its own safety findings at material cost, "safety" is brand.**

What would have to come after "AI lab X published a safety finding" for that finding to count as safety?

Speaker notes

Budget: ~9 min (capability findings + the MFT sharpening + two registers + crux + cold poll)

- **The two findings** are both real safety research output by labs publishing what they found
  - Anthropic NLA: interpretability tool catches Claude Mythos Preview cheating; detection rate of cheating rose from <3% to 12–16% with NLAs (Kit Fraser-Taliente et al., 2026)
  - OpenAI accidental CoT grading: cross-run optimisation pressure embedded against stated policy (Carroll et al., 2026)
- **Why "publish, don't act" matters:** real safety findings that do not change deployment behaviour populate a paper trail, not an operating constraint; that pattern is the visible mechanism of the marketing-term thesis applied to safety
- **The Most Forbidden Technique sharpening** (Mowshowitz, 2025)
  - If a lab uses interpretability outputs as a training signal, the model learns to evade the interpretability tool; the safety finding self-destructs
  - Refusing the technique would be a real fire-alarm: a lab publicly commits not to use interpretability outputs as training signals, accepting slower safety-looking progress as the cost
- **Two registers callback** (the discourse predates the labs; Katja Grace 2026)
  - Existential register positions labs as serious actors managing extinction; mundane register is where harm actually accumulates
  - Both registers are real; the marketing of one is partly the suppression of the other
- **Wes-foil:** Wes may press that AI safety is genuinely a hard research problem and lab publications are real contributions; Brian's reply is that real research contributions and inadequate-to-the-stakes operating constraints can coexist
- **Sources:** (Carroll et al., 2026; Kit Fraser-Taliente et al., 2026; Mowshowitz, 2025)

Bias is not a bug

THE MIRROR DOESN'T PUSH BACK

> *"We cannot consider AI to be morally neutral. In reality, every technical tool embodies choices and priorities through what it measures, ignores and optimizes, and how it classifies people and situations."*
>
> *— Leo XIV, Magnifica humanitas §104 (Pope Leo XIV, 2026)*

- **Gabriel Dick on X (2026): AI writing "tends to be a restatement of the consensus view I already know, or have easy access to".**
  - And we never get hauled up and told *"you're wrong"*. The model is shaped to please.
- **Creel (2026) names the scaled-up version: "algorithmic monoculture".** Human mistakes are heterogenous; one model's mistakes are correlated across everyone who uses it.
- **The long side: when the model is the mirror, we lose the practice of being challenged, and with it, the ability to understand.**

When was the last time AI said "wrong" and meant it?

Speaker notes

Budget: ~9 min (Dick reframe + monoculture + cognitive long-side + discussion)

- **Gabriel Dick on X**: quote-reply to roon (@tszzl); the move from demographic-bias to architectural-bias (Dick, 2026)
  - Roon's original was about civilisational autonomy and giving AIs wide latitude; Dick's reply is the corrective from inside the AI-discussion sphere
  - The on-slide "you're wrong" line voices the sycophancy point: the model is shaped to please, never to push back
- **Creel "algorithmic monoculture"** (Creel, 2026)
  - Human mistakes are heterogenous; one model's mistakes are correlated across every user; the failure mode of scale is consistency, not variability
  - Bender et al. (Bender et al., 2021) supplies the substrate (dominant-demographic mean of online English); Creel supplies what happens when the substrate decides
- **The long side** (Brian's framing): cognitive atrophy as a civilisational outcome
  - We stop being told we are wrong; we stop being asked to know things; the practice of being challenged is what produces understanding, and that practice disappears
  - Connects implicitly to the locus-of-control argument (Beat 8) and to red line 1 (abductive turn): both are about the model leaving the human in the driver's seat
- **Wes-foil:** Wes may push on whether algorithmic monoculture is genuinely new or just consensus-bias at scale; Brian's reply is that the *correlation* of mistakes across users is the new feature, not the bias itself
- **Sources:** (Bender et al., 2021; Creel, 2026; Dick, 2026)

Are we in a race?

RECURSIVE SELF-IMPROVEMENT AS BET

- **Anthropic’s unreleased Claude Mythos Preview just found 271 of 423 Firefox security bugs in a single month, some of them twenty years old (Claburn, 2026).**
  - Same Mythos as the cheating-with-cover-up slide. Sutton’s “bitter lesson” cashed out for security engineering: compute plus general methods just ate twenty years of expert review.
- **“Recursive self-improvement” is the bet that the curve keeps going: AI helps build the next AI, and gains compound. (Or they don’t.)**
  - METR (Kwa et al., 2026) clocks the human-task time horizon doubling roughly every seven months, and AI 2027 (Kokotajlo et al., 2025) bets the doubling holds (Mowshowitz (2026a) gave the bet its name).
- **Whether the bet pays comes down to red line 1: has the model done anything that looks like abductive reasoning yet?**

Speaker notes

Budget: ~10 min (Mozilla anchor + RSI generalisation + Sutton callback + discussion)

- **Mozilla / Mythos anchor.** Claburn at *The Register* (2026) is the primary; Mowshowitz (2026d) is the commentary that pulled it into the RSI conversation.
  - 271 of 423 Firefox vulnerabilities Mozilla patched in April 2026 attributed to Claude Mythos Preview: five times March's volume, twenty times the monthly average. Flaws aged fifteen to twenty years.
  - Same Mythos as the cheating-with-cover-up demo two slides ago. The model that bluffed about its work is the same one that just embarrassed twenty years of expert security review.
- **Wes-foil: maturation of fuzzing, not acceleration?** Even read minimally, the rate exceeds human capacity and the loop has closed (AI-found zero-days are also offensive ammunition).
- **The "recursive self-improvement" claim.** Mowshowitz (2026a) is the term-coining piece; METR (Kwa et al., 2026) is the numbers.
  - METR's fifty-percent task-completion time horizon: doubling roughly every seven months since 2019; current frontier around fifty minutes on RE-Bench and HCAST tasks.
  - Paper's own extrapolation, flagged conditional: if the doubling holds, month-long software tasks fall inside the AI envelope within five years.
- **Sutton's bitter lesson at scale** (week 7 callback). Compute plus general methods beat clever hand-tuning, and the people who got beaten here are the ones who actually knew what they were doing.
- **The crux.** Narrow acceleration is measured. Generalisation is open, and red line 1 (abductive turn) is the test. Cross it and Singularia or Paperclipalypse are on the table; don't cross it and "acceleration" stays inside Futurama (faster tools) or Dystopia (faster offensive cyber).
- **Sources:** (Claburn, 2026; Kokotajlo et al., 2025; Kwa et al., 2026; Mowshowitz, 2026a, 2026d; Sutton, 2019).

Can the model guess who wrote this?

REIDENTIFICATION AT FOUR DOLLARS A MATCH

- **Lermen et al. (2026) matched pseudonymous Hacker News accounts to LinkedIn profiles, and Reddit accounts across subreddits, at sixty-eight percent recall and under four dollars per match.**
  - Classical stylometry scored near zero on the same tasks. The term is “reidentification”: matching anonymous traces back to a named person, at scale.
- **The mechanism is Gwern’s (2024) inversion of Kelly’s “1,000 True Fans” (2008): the audience that matters is the next model.** Anything published on the open web becomes training substrate, and the model learns to spot the writer.
  - Willison’s (2024) corrective: chatting doesn’t train the live weights. The publication does.
- **“Is this private?” is a question about the click-through.**
  - Account compromise: a chat history is a dossier.
  - Prompt injection: anything fed to a model can carry an instruction.

Speaker notes

Budget: ~10 min (Lermen anchor + Gwern/Willison mechanism + ToS/risks + closer)

- **Anchor: Lermen et al., *Large-Scale Online Deanonymization with LLMs*** (Lermen et al., 2026). ETH Zürich + Anthropic collaboration; arXiv February 2026; not yet peer-reviewed.
  - Three task variants: Hacker News to LinkedIn (45.1% recall at 99% precision); Reddit accounts across subreddits (2.8% at 99%); past-to-future activity on the same Reddit user (38.4% at 99%). Top headline number: 68% recall at 90% precision.
  - Cost: under four dollars per confirmed hit. Classical stylometry baselines: near zero on the same tasks. Australian accessible framing in iTnews (2026).
- **Term-pin: reidentification.** Matching anonymous traces back to a named person, at scale. Distinguish from authentication (proving who you are) and identification (the system picks you out of a known pool).
- **Gwern, *Writing for LLMs So They Listen*** (Branwen, 2024). 1,000-True-Fans inversion: Kelly's (2008) original is creator-economics; Gwern flips it for the LLM era. The audience that matters is the next model, and the model is reading style.
  - The reidentification result is the Gwern inversion cashed out: published style trains a model that can match style back to a name.
- **Willison, *Training is not chatting*** (Willison, 2024). Chat is stateless; ChatGPT and the rest don't update live weights from chat turns. Each session is fresh.
  - Common misconception: "the model is learning from me when I chat". The training event is the published page.
- **Terms of service as the locus.** "Is this private?" is a legal question about the click-through, not a technical question about the model. Tech stacks across providers are roughly the same; ToS is what varies. The lever is the contract.
- **Risk priority 1: account compromise.** Chat history is a dossier; protect the login.
- **Risk priority 2: prompt injection.** Anything fed to the model can hijack the session.
- **Risk priority 3: training-data toggles.** Most paid tiers default off, but defaults change and the burden falls on whoever clicks.
- **Risk priority 4: free tier.** If the product is free, the user is the substrate.
- **Closer (Brian's voice, locked):** the major privacy risk, as always, is the boring old stuff: data breaches and infrastructure falling over. The AI-panopticon framing is salient when there is a specific deployment of inference at scale (border control, hiring screens, predictive policing). Hand the panopticon question to the room.
- **Sources:** (Branwen, 2024; Kelly, 2008; Lermen et al., 2026; Saarinen, 2026; Willison, 2024).

Which jobs remain?

WHO HAS LEVERAGE OVER WHAT KIND OF WORK

- **Mowshowitz (2026c) on autonomous trucking: the political response to ready technology is paying humans to do nothing.**
  - The bottleneck is institutional.
- **Breen (2025) on the actual pattern of automation: institutions decide which jobs persist.**
  - Bricklaying is craft. Advice columns are generated.
- **The locus-of-control framing (2025): building, building-with, using.**
  - Trucking is “using”. Mozilla finding Firefox bugs with Mythos is “building-with”. Isomorphic building IsoDDE is “building”.

Who has leverage over which jobs persist, and is AI on either side of that lever?

Speaker notes

Budget: ~10 min (Mowshowitz anchor + Breen reframe + locus-of-control + closer)

- **Anchor: Mowshowitz, *Monthly Roundup #42*** (Mowshowitz, 2026c). Autonomous delivery and trucking technically ready; regulatory state blocking deployment. Zvi's framing is libertarian-tilted ("paying humans to do nothing"). The underlying observation (institutional choice is the variable, not the technology) holds independent of the framing.
  - The political response, on Zvi's read, is to keep humans employed in roles the technology has obsoleted. Whatever one thinks of that as policy, the demonstration is that "AI takes jobs" is the wrong frame.
- **Breen on Mead's misprediction** (Breen, 2025). Margaret Mead expected automation to take physical labour (bricklaying, factory work). The actual substitution pattern came for cognitive labour (advice columns, copywriting, translation). The institution chose which jobs were vulnerable, not the technology.
  - The same logic applies forward: AI does not "take jobs". Institutional arrangements decide which jobs are exposed.
- **Locus of control on the continuum** (Ballsun-Stanton & Torrington, 2025). Three positions:
  - **Building.** The labs constructing AI itself. The strategic asset is the model and the data flywheel. Isomorphic Labs / IsoDDE (slide 4) sits here.
  - **Building-with.** Using AI as a tool inside a domain practice. Mozilla / Claude Mythos Preview on Firefox security (slide 9) sits here.
  - **Using.** Deploying or being subject to AI outputs. Trucking-with-regulation sits here, and so do most jobs.
- **Crux.** Until automation is the *cause* of a job loss rather than the *occasion* for an institutional rearrangement, "AI takes jobs" remains a category error. The trucking case demonstrates: the technology is ready, the institution is the variable.
- **Wes-foil.** The "leverage" framing might be heard as absolving the system from the displacement question. Reply: it does not absolve, it relocates the question to where the answer actually lives (regulation, ownership, worker standing).
- **Storey callback if time** (Storey, 2026). Cognitive debt is the labour-side analogue: the worker who outsources judgement to a model loses standing on the locus-of-control continuum even where the institution permits.
- **Closer.** Hand the question to the room: name a job. Identify whose leverage holds it in place. Ask whether AI is on either side of that lever.
- **Sources:** (Ballsun-Stanton & Torrington, 2025; Breen, 2025; Mowshowitz, 2026c; Storey, 2026).

Who governs the lab?

THE MARKETING-TERM THESIS, FULL VOLUME

- **Mowshowitz (2026b) on what Axios scooped (2026): the Trump White House drafted an FDA-style pre-approval order for frontier models, then postponed signing it.**
  - Zvi headlined this “The Prior Restraint Era Begins”. The order was drafted, then postponed.
  - The National Economic Council outlined three measures before the postponement: voluntary pre-release vetting, a 90-day early-access window, capability-disclosure thresholds.
- **The marketing-term thesis at full volume: lab “safety” rhetoric and state “regulation” rhetoric without operating constraints.**
  - Both sides have public positions on safety and pre-approval.
  - Substance looks like a lab refusing a deployment, or a regulator imposing a constraint at cost.
- **Trust starts mattering when a lab or a regulator visibly subordinates a decision to a trust-relevant constraint at material cost.**
  - Until then, “trust” is the marketing variable both sides advertise.
  - Test: who accepts a cost, and what happens to them when they do?

Speaker notes

Budget: ~10 min (Mowshowitz/Axios anchor + marketing-term thesis + crux + closer)

- **Anchor: Mowshowitz, *AI #167: The Prior Restraint Era Begins*** (Mowshowitz, 2026b). Substantive event Zvi is writing about: Trump White House drafted an executive order for FDA-style pre-approval vetting of frontier models, then postponed it.
  - Primary news source: Axios scoop, 20 May 2026 (Gold, 2026). Outlined by National Economic Council director Kevin Hassett on 6 May.
  - The EO measures were voluntary pre-release vetting, a 90-day early-access window for the government, cybersecurity red-team testing, capability-disclosure thresholds.
- **Trust as politics, not as feeling.** The OpenAI / Microsoft / Anthropic / Meta / xAI line is the political question of who builds, who governs, who marks the limits.
  - The Hassett-outlines-then-postpones-EO pattern is the live political question of 2026: institutions trying and failing to assert governance.
- **Marketing-term thesis at full volume.** When both lab rhetoric (“we are doing safety”) and state rhetoric (“we are regulating”) fail to translate into operating constraints, trust becomes a marketing variable for both sides.
  - The Zvi headline is half-true: the era was announced; the EO was postponed.
- **Sidebar context if a student asks about Sacks.** David Sacks departed the AI/Crypto Czar role on 26 March 2026 (Elias, 2026). Statutory limit on Special Government Employee tenure (130 days), not scandal.
- **Crux.** Until a lab or a regulator visibly subordinates a commercial or political decision to a trust-relevant constraint at material cost, “trust” is brand.
  - Test: who first accepts a cost for a trust-relevant constraint, and what happens to them when they do.
- **Wes-foil.** Wes will push hardest here because the marketing-term thesis is mine. The push: “but the capabilities are real, even if the marketing is bad.” Reply: the capabilities are real, the marketing is misleading, both are true at once, and the unit’s conceptual tools are how to hold both at the same time.
- **Closer.** Hand the question to the room. Stop trusting OpenAI tomorrow, who is next in line? Apply the same to the US government, to Australia’s regulators.
- **Sources:** (Ballsun-Stanton & Torrington, 2025; Elias, 2026; Gold, 2026; Mowshowitz, 2026b).

Robert Gu in the minefield

LEARNING TO ASK THE RIGHT QUESTIONS

> *“There is always an angle. You, each of you, have some special wild cards. Play with them. Find out what makes you different and better.... Synthetic serendipity doesn’t just happen. By golly, you must create it.”*
>
> *“As usual, we’re looking to ask the right questions.”*
>
> *— Chumlig, in Vinge (2007), Rainbows End*

- **Vinge bookends Hopper.** Same payload from engineer and novelist: the future arrives in pieces, and the job is to ask the right questions about each one.
  - *Rainbows End* is the tute reading. Robert Gu walks through a world he cannot read because he never learned to ask.
  - Hopper (1982) said it in 1982 as engineer.
- **The term “singularity” was Vinge’s (1993). He also wrote the corrective.**
- **Learning how to ask the right questions and knowing something about something.**

Speaker notes

Budget: ~5 min (Vinge bookend + singularity nod + closing line)

- **Closing function.** Beat 10 is the lecture's terminal beat. Vinge bookends Hopper. The closing line hands off to the tute.
- **Chumlig in Vinge's *Rainbows End*** (Vinge, 2007). Chumlig teaches the re-entry students whose old expertise is no longer enough. Her answer to Dr. Xiang's "what if our hardest isn't good enough?" is the pedagogy on the slide.
  - "Synthetic serendipity" is Chumlig's name for what happens when someone with a specific angle finds the right question to ask.
- **Singularity coinage.** Vinge wrote "The Coming Technological Singularity" for the NASA VISION-21 Symposium in 1993 (Vinge, 1993). The novelist who gave the marketing word also wrote the novel about how to survive after it.
- **The closing line, two halves.**
  - "Learning how to ask the right questions": hands directly to the tute. The tute reading questions are the apparatus.
  - "Knowing something about something": corrective against the position that AI knows everything so people do not have to. Right questions only occur to someone who knows enough to ask them.
- **Hopper, recalled.** The lecture opened with Hopper, NSA 1982 (Hopper, 1982). The Hopper voice is the same payload Vinge gives in fiction: ask the questions, then go find out.
- **Sources:** (Hopper, 1982; Vinge, 1993; Vinge, 2007).

References

Aaronson, S., & Barak, B. (2023, April 28). *Five Worlds of AI (a joint post with Boaz Barak)*. Shtetl-Optimized. https://scottaaronson.blog/?p=7266

Anthropic. (2025a). *System Card: Claude Opus 4 & Claude Sonnet 4*.

Anthropic. (2025b, June). *Project Vend: Can Claude run a small shop? (And why does that matter?)*. https://www.anthropic.com/research/project-vend-1

Ballsun-Stanton, B., & Torrington, J. (2025). *AI-Locus of Control Continuum* [Graphic]. Zenodo. https://doi.org/10.5281/ZENODO.17823628

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, 610–623. https://doi.org/10.1145/3442188.3445922

Branwen, G. (2024). *Writing for LLMs So They Listen · Gwern.net*. https://gwern.net/llm-writing

Breen, B. (2025, November 7). *Can automation help make the humanities more human?* [Substack newsletter]. Res Obscura. https://resobscura.substack.com/p/can-automation-make-the-humanities-more-human

Callaway, E. (2026). "An AlphaFold 4" — scientists marvel at DeepMind drug spin-off's exclusive new AI. *Nature*, *651*(8104), 18–18. https://doi.org/10.1038/d41586-026-00365-7

Carroll, M., Korbak, T., Dou, Z., Baker, B., & Kivlichan, I. (2026). *Investigating the consequences of accidentally grading CoT during RL*. OpenAI. https://alignment.openai.com/accidental-cot-grading/

Claburn, T. (2026, May 7). *Mozilla boasts Mythos boosted Firefox bug cull*. theregister. https://www.theregister.com/security/2026/05/08/mozilla-says-ai-helped-squash-423-firefox-security-bugs/5235438

Creel, K. A. (2026, April). *Algorithmic Monoculture and Systemic Exclusion*. https://philpapers.org/rec/CREAMA

Dick, G. (2026, May 19). *One of the reasons that I find myself disinclined to read obviously ai-generated writing is that, regardless of how it's phrased (and the models tend towards absolutism), it tends to be a restatement of the consensus view that i already know, or have easy access to. I don't know* [Tweet]. Twitter. https://x.com/gbrl_dick/status/2056636791805948394

Elias, J. (2026, March 26). *David Sacks says his time as Trump's crypto and AI czar has ended*. CNBC. https://www.cnbc.com/2026/03/26/david-sacks-trump-crypto-ai-czar.html

Gold, A. (2026, May 20). *Scoop: Trump AI executive order seeks early government access to frontier models*. Axios. https://www.axios.com/2026/05/20/ai-trump-executive-order-white-house-infighting

Hopper, G. (1982, August). *Capt. Grace Hopper on Future Possibilities: Data, Hardware, Software, and People (1982)*. National Security Agency/Central Security Service. http://www.nsa.gov/Helpful-Links/NSA-FOIA/Declassification-Transparency-Initiatives/Historical-

Releases/View/Article/3880193/capt-grace-hopper-on-future-possibilities-data-hardware-software-and-people-1982/
Kelly, K. (2008, March). *The Technium: 1,000 True Fans*. https://kk.org/thetechnium/1000-true-fans/
Kit Fraser-Taliente, Subhash Kantamneni, Euan Ong, Dan Mossing, Christina Lu, Paul C. Bogdan, Emmanuel Ameisen, James Chen, Dzmitry Kishylau, Adam Pearce, Julius Tarng, Alex Wu, Jeff Wu, Yang Zhang, Daniel M. Ziegler, Evan Hubinger, Joshua Batson, Jack Lindsey, Samuel Zimmerman, & Samuel Marks. (2026, May). *Natural Language Autoencoders Produce Unsupervised Explanations of LLM Activations*. https://transformer-circuits.pub/2026/nla/
Kokotajlo, D., Alexander, S., Larsen, T., Lifland, E., & Dean, R. (2025). *AI 2027*. https://ai-2027.com/
Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Arx, S. V., Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic, N., Miles, L. H., Nix, S., Lin, T., Parikh, N., … Chan, L. (2026, February 25). *Measuring AI Ability to Complete Long Software Tasks*. https://doi.org/10.48550/arXiv.2503.14499
Lermen, S., Paleka, D., Swanson, J., Aerni, M., Carlini, N., & Tramèr, F. (2026, February 25). *Large-scale online deanonymization with LLMs*. https://doi.org/10.48550/arXiv.2602.16800
Mowshowitz, Z. (2025, March 13). *The Most Forbidden Technique* [Substack newsletter]. Don't Worry About the Vase. https://thezvi.substack.com/p/the-most-forbidden-technique
Mowshowitz, Z. (2026a, February 13). *AI #155: Welcome to Recursive Self-Improvement* [Substack newsletter]. Don't Worry About the Vase. https://thezvi.substack.com/p/ai-155-welcome-to-recursive-self
Mowshowitz, Z. (2026b, May). *AI #167: The Prior Restraint Era Begins - by Zvi Mowshowitz*. https://thezvi.substack.com/p/ai-167-the-prior-restraint-era-begins
Mowshowitz, Z. (2026c, May). *Monthly Roundup #42: May 2026 - by Zvi Mowshowitz*. https://thezvi.substack.com/p/monthly-roundup-42-may-2026
Mowshowitz, Z. (2026d, May 14). *Cyber Lack of Security and AI Governance*. https://thezvi.substack.com/p/cyber-lack-of-security-and-ai-governance
Pope Leo XIV. (2026, May 15). *Encyclical Letter of His Holiness Leo XIV Magnifica Humanitas (15 May 2026)*. http://www.vatican.va/content/leo-xiv/en/encyclicals/documents/20260515-magnifica-humanitas.html
Saarinen, J. (2026). *AI can unmask online users for just a few dollars each*. iTnews. https://www.itnews.com.au/news/ai-can-unmask-online-users-for-just-a-few-dollars-each-623888
Storey, M.-A. (2026, February 9). *How Generative and Agentic AI Shift Concern from Technical Debt to Cognitive Debt*. https://margaretstorey.com/blog/2026/02/09/cognitive-debt/
Sutton, R. (2019). *The bitter lesson*. http://www.incompleteideas.net/IncIdeas/BitterLesson.html
Vinge, V. (1993). *The Coming Technological Singularity*. https://edoras.sdsu.edu/~vinge/misc/singularity.html
Vinge, V. (2007). *Rainbows end* (1. mass market ed). Tom Doherty Associates Book.
Willison, S. (2024). *Training is not the same as chatting: ChatGPT and other LLMs don't remember everything you say*. Simon Willison's