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)*

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?

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?

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?

Performance of evil vs actual evil

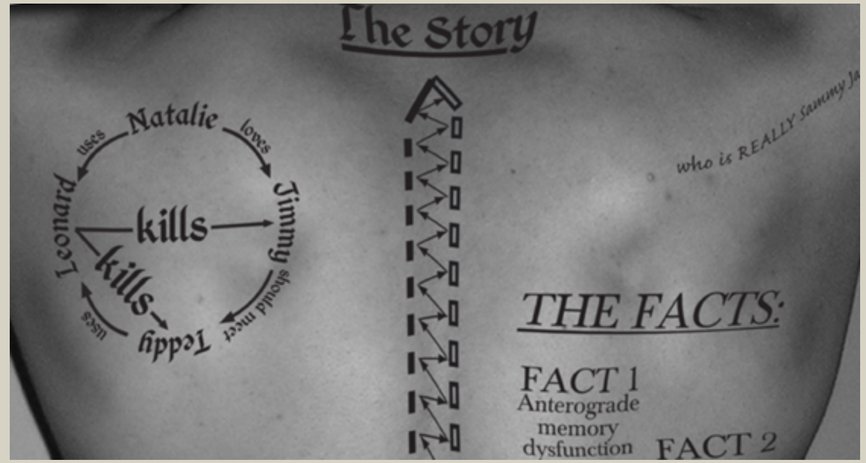
RED LINE 2: WHO DID THE BLUFFING?

- **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.

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*

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?

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?

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?**

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.

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?

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?

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.**

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