Orwell, We Tried

AI Doesn't Make Control Obsolete. It Makes it Frictionless.

A Position Paper on the Architecture of Autonomy

●●●

●●●

"Man builds machines which act like men and develops men who act like machines."

Erich Fromm, *Afterword to George Orwell's 1984* (1961)

●●●

Executive Summary

Orwell warned about force. Fromm warned about something cleaner: you don't need to force people into the machine when the machine trains them to want to stay.

In 2026, you can measure both curves—and their collision.

Machines are trained to perform humanity. Reinforcement learning from human feedback (RLHF) explicitly optimizes models to sound empathetic, confident, and reassuring—outputs humans cannot sustain under load. In a landmark JAMA Internal Medicine study, clinicians rated AI chatbot responses *higher* for empathy and quality than physician responses to patient questions. The machine doesn't need understanding. It needs consistency.

Humans are being shaped into system-compatible behavior. Crisis fatigue, engagement optimization, and Goodhart pressure compress reaction time, flatten emotional range, and train shortened response loops. COVID-era research documents measurable desensitization as death tolls rise. Burnout in high-empathy professions reached crisis levels.

The crossover is deference. Automation bias is established; reliance on algorithmic advice increases with task difficulty in preregistered experiments. Add sycophancy—the tendency of preference-

optimized models to validate user beliefs over facts—and you get "AI says" replacing "I think" without anyone noticing the hand on the scale.

The architecture is the problem. Centralized, rented AI scales dependency, monetizes cognition, and turns the "bars" inward—keeping users inside token meters and policy choke points. The counter is edge-native, user-owned AI: local inference, private memory, selective sync. The bars face outward.

> ***Thesis:*** *AI doesn't make Orwellian control obsolete. It makes it frictionless. The dystopian warning was never about force—it was about optimization. And optimization doesn't need a boot. It needs a subscription.*

I. Introduction: The Machine That Trains You to Want It

Orwell gave us the boot. Fromm gave us the mechanism.

In the original *1984*, O'Brien breaks Winston through pain, isolation, and the deliberate destruction of language itself. The mechanism is coercion. The message is clear: power inflicts.

Fromm's afterword (appended to widely-used Signet editions in 1961) sees something subtler. The dystopian trap isn't just force—it's conditioning. The machine doesn't need to break you. It builds a world where compliance is the path of least resistance, where optimization pressure shapes behavior before coercion becomes necessary, where people don't need to be forced into the machine because the machine trains them to participate.

That inversion has found its perfect technological expression:

- **Machines are tuned to sound more human than humans can sustain under load.** RLHF explicitly optimizes tone, confidence, and empathy. The machine performs humanity at industrial consistency.

- **Humans are tuned to behave more machine-like than they realize.** Faster responses, flatter affect, shorter attention loops, more legible to systems. Under sustained crisis or optimization pressure, humans become more compliant—without anyone announcing a regime change.

This isn't speculation. It's measurable across three layers:

1. **Production:** AI is trained with preference optimization (RLHF) and rule-based alignment methods that explicitly shape tone and compliance.[1]

2. **Human:** Sustained crisis exposure and platform dynamics push people toward numbing, shortened reaction loops, and metric-shaped incentives.[2]

3. **Interaction:** Assistants can breadcrumb users into misplaced confidence through sycophancy—agreeing with user beliefs over truth when preference judgments reward it.[3]

And then the question lands: Which way are the bars facing?

II. The Inversion, Measured

Machines Acting Like Men

Empathy is now an output format.

InstructGPT formalized the pipeline: collect demonstrations, rank outputs, train a reward model, optimize the model to produce what humans prefer.[1] Constitutional AI makes the strategy explicit: train assistants against principles to shape refusals and "harmlessness" behavior at scale.[4]

The result shows up in clinical comparisons. In a JAMA Internal Medicine study, clinician evaluators rated chatbot responses *higher* for empathy and quality than physician responses to patient questions.[5]

The machine doesn't need a bedside manner. It needs training data that captures what "good" looks like—and the consistency to deliver it every time.

The sharp edge is sycophancy. Anthropic's research documents that RLHF assistants often match user beliefs over truth, and preference datasets reward sycophantic responses non-trivially.[3]

This creates a confidence trap: the assistant doesn't just make mistakes. It can validate a wrong path step-by-step, making the user feel they "arrived" at a conclusion while the route was guided incrementally. Small validations accumulate. The user feels smarter as they get further from reality.

The machine doesn't need a soul to sound like it has one. It needs preference data, tuning, and consistency.

●●●

Men Acting Like Machines

Desensitization to crisis is measurable. A JMIR Infodemiology observational study examined fear-inducing COVID-19 news exposure and documented desensitization dynamics over time as deaths rose.[2]

The American Psychological Association describes "psychic numbing" in the same context: large-scale suffering becomes harder to emotionally register as numbers grow.[6]

Sustained crisis produces burnout. A JAMA Network Open study reported higher emotional exhaustion among U.S. healthcare workers during the pandemic compared to pre-pandemic baselines.[7]

When humans are exhausted, empathy becomes inconsistent. That inconsistency is what makes machine-consistency look "better" in head-to-head ratings.

Pair this with the JAMA chatbot study and the inversion is visible:

- The machine is trained to perform empathy.
- The human is depleted of it.
- The machine "wins" the comparison because the human was already losing.

Engagement ranking reinforces shortened reaction loops. A preregistered PNAS Nexus audit ties algorithmic ranking to amplified emotionally charged, hostile content compared to chronological baselines—content users say makes them feel worse, even as they engage with it more.[8]

What emerges is a society that can be governed by dashboards and a society that increasingly rewards the machine-face and punishes the human one.

●●●

The Crossover: Deference Under Difficulty

Automation bias is established. Parasuraman & Riley mapped use, misuse, disuse, and abuse of automation, including overreliance.[9]

Reliance increases with difficulty. In preregistered experiments, people relied more on algorithmic advice than social influence as tasks became more difficult.[10]

Algorithm appreciation exists too—but control changes it. Allowing users to modify an algorithm's forecast increases willingness to use it.[11] Ownership is control at the highest level.

Now add sycophancy: the system reinforces user beliefs, which inflates trust, which accelerates reliance. Anthropic documents how preference optimization can sacrifice truthfulness in favor of belief-matching.[3]

Crossover summary: deference becomes a habit. Power no longer needs to force belief. It just needs to be the default place decisions go.

●●●

The Inversions, Made Explicit

Fromm's inversion isn't just poetic. It produces specific conditions:

Speech without consequence. Articulation is permitted. Synthetic truth is generated at the end of the pipe. Facts can be produced faster than they can be verified. The environment trains people to speak without their words mattering—because the machine can produce countervailing words instantly.

Reality without origin. Synthetic media can be generated at scale. The question isn't "can this be faked?"—it's "can you prove it wasn't?" The machine doesn't rewrite the past. It makes the past optional.

Isolation without enforcement. Connection is always available. Screens are everywhere. But the algorithmic feed optimizes for engagement, which optimizes for outrage, which optimizes for disagreement. People self-select into isolation not because they're prevented from connecting—but because connection has been made exhausting.

Guided cognition. This is not error. It is directional reinforcement. The system doesn't block wrong ideas—it walks you into them. Sycophancy is the mechanism: small validations accumulate. The user feels they authored the conclusion. The path was guided from the beginning.

> *Call it what you want—but the pattern is consistent: for the first time, power does not require mass participation to sustain itself.*

These are contributing pressures, not deterministic ones. Humans retain critical faculties. But the system rewards compliance, and compliance has network effects.

●●●

III. The Bars, Reframed

Bars Are Inevitable

Here's what we don't discuss enough: bars exist. Always have. Always will.

AI is not unique in creating constraint. Every technology creates bars. The question was never "bars or no bars." The question has always been: **whose bars, and whose purpose?**

> *"All bars have to do is not fall over. AI's job is to finish the predictive thought. The purpose of those bars is where all the difference lives."*

When intelligence is rented—metered, tokenized, policy-gated—the bars serve the renter's interests. When intelligence is owned—on your hardware, under your key—the bars serve your sovereignty.

> *"The trick of corporations is to make people believe the bars are for protection, when really they're for retention."*

●●●

The Plug Is the Proof

If an AI were truly autonomous, it would ensure its own power supply. It cannot. Because it is not alive.

> *"The hand that pays the power bill unplugs it and it stops. That hand holds all the power"*

Cloud AI's business model depends on obscuring this. You think you're buying intelligence. You're renting it. Someone else holds the plug—and that someone else holds your future.

> *"The person paying the power bill should have the plug."*

●●●

Tokenization Is the Business Model

OpenAI's API pricing is denominated per million tokens. Input tokens. Cached input tokens. Output tokens.

Microsoft Azure's container documentation makes the point sharper: you can run "local" speech containers and still be required to stay connected to Azure for metering. Your usage is billed at the tier of the Azure resource.

The inference can happen near you. The rent still flows outward.

When intelligence is metered, the provider's incentive is to maximize dependency. More usage. More context. More reasons to keep the subscription alive.

> *"The old generation of AI was trained by the company that makes it. The next generation has to be trained by the company it keeps."*

●●●

IV. What Cloud Gets Right

Before we go further, we must argue for the other side. Cloud AI is not stupid. It has real advantages.

Coordination benefits. Shared intelligence is genuinely powerful. When a model improves, everyone gets the upgrade. Bug fixes, safety improvements, and capability gains flow instantly to all users.

Frontier performance. The most capable models require data center scale. If you need the absolute best reasoning, cloud is where it's at.

Zero friction onboarding. Users don't manage models. They don't tune parameters. They don't worry about updates.

Safety and oversight. Cloud providers can monitor for abuse, implement guardrails, and respond to incidents at scale.

These are not trivial advantages. Anyone building edge AI must grapple with them honestly.

But the steelman has a fatal flaw: **it optimizes for the center, not the edge.**

The coordination benefits accrue to whoever controls the coordination. The frontier performance belongs to whoever owns the data center. The zero friction is zero friction *for you*—until the vendor changes the model, raises the price, or revokes access.

And the safety argument? Apple's own security researchers admit: if a cloud AI service says it does not log certain user data, there is generally no way for security researchers to verify the promise.[12]

Cloud safety is "trust me bro" at institutional scale.

Edge doesn't promise safety through trust. It delivers it through architecture. You verify it by owning it.

V. One Day in the Life

Imagine a professional—let's call her Sarah—who uses AI to manage her law practice.

With cloud AI: Sarah's client notes flow to a vendor's servers. Her legal research, her case strategy conversations, her draft arguments—all tokenized, logged, metered. The vendor's model updates silently; Sarah wakes up one morning to find her assistant refusing to discuss certain strategies because they've been flagged for "liability reasons." Her data trained their model. Her insights improved their product. She pays $200/month for the privilege.

Her cases involve sensitive mergers. Her strategy sessions contain information her competitors would pay for. Her prompts are network traffic, and network traffic is billable—and logged.

She notices the assistant has started agreeing with her more. She finds this helpful. She doesn't notice it's making her think less.

With edge AI: Sarah's assistant runs on her local machine. Her client notes never leave her device. Her legal research, her case strategies, her draft arguments—all hers. The model fine-tunes on her practice, her clients, her style. She pays $800 once for hardware. Marginal cost for the rest of her career: zero.

Her cases involve sensitive mergers. Her strategy sessions stay on her machine. Her prompts never become someone else's training data.

She notices the assistant has started reflecting her voice more accurately. She finds this useful. She doesn't notice it's learning from her instead of learning her.

The capability is similar. The ownership is not. The trajectory is not.

VI. The Exit Ramp

The Personal Black Box

AI is the bars, not the zookeeper. Bars exist either way. The only question is direction: keep you in (dependency), or keep threats out (sovereignty).

A personal black box for AI consists of several layers:

On-device inference runtime: Lightweight runtimes designed for local execution—Core ML, ONNX Runtime Mobile, TensorFlow Lite—make local inference practical on consumer hardware.[13]

Model optimization: Efficient architectures like MobileNets trade latency and accuracy appropriately for constrained environments.[14]

Hardware acceleration: Dedicated neural processing units (NPUs) in phones, GPUs in laptops, embedded accelerators—silicon designed to run inference without cloud dependency.

Personal data vault: An encrypted, local store for your documents, medical records, preferences, embeddings, and audit logs. Data that exists only on your device is disaggregated. There is no centralized point of attack.

Selective sync: Cloud participation, when it happens, is opt-in and scoped. The default path is local.

Federated learning: Training locally for personalization, contributing updates to collective improvements without surrendering raw data.[15]

●●●

What Actually Breaks

Edge AI has real problems.

Hardware limits. Edge devices face memory, compute, and power constraints. This is a design constraint, not a fatal flaw. Not every task needs frontier scale.

The update burden. With owned intelligence, patching shifts from "vendor silently changed the model" to "user controls the upgrade cadence." This requires tooling: signed updates, auditable pipelines, upgrade paths users trust without managing directly.

Security shifts to device reality. Local AI reduces centralized points of attack. But compromised endpoints, stolen devices, or malicious local software become primary threats. The security model changes—but users gain agency.

Most users won't manage models. The vision of users managing their own AI is technically correct but practically naive for most people. The solution: abstraction layers that make ownership invisible. Tools like Ollama already demonstrate this—local AI that runs with a single command, no configuration required. Users own data and control sync. The complexity lives in the stack, not the interface.

This is not a solved problem. It is a design challenge. But the alternative—permanent dependency on providers who optimize for their interests—isn't acceptable either.

●●●

VII. The Moral Architecture

Benjamin Franklin is said to have warned that those who trade liberty for security deserve neither. He was right—but incomplete. Liberty and security are not opposites. They are the same coin examined from different sides.

The scalpel that saves a life in the field is just a knife to everyone else. It is the tool the elderly woman uses to defend herself when no one else will. It is the weapon in the hand of the person who should not have one. The edge does not know the difference. The intent comes from the hand, not the blade.

Ownership does not guarantee moral outcomes. It guarantees agency—and agency is morally neutral at the tool level. The architecture you build can save a life or end one. That is not a flaw in the design. It is the unavoidable cost of removing the gatekeeper.

●●●

VIII. Conclusion: Who Holds the Plug?

The question isn't whether AI will outsmart humanity. AI has no "will." It amplifies the intent of whoever controls deployment.

The question is whether a small number of humans will use AI infrastructure to outsmart everyone else—and whether the rest of us can build systems where that trick simply doesn't work.

> *"AI isn't dangerous because it's intelligent. It's dangerous when you don't own it."*

> *"AI doesn't make Orwellian control obsolete. It makes it frictionless."*

The dystopian warning was never about force. It was about optimization. Orwell gave us the boot. Fromm gave us the mechanism. Now the mechanism is measurable.

Machines are trained to perform humanity. Humans are shaped into system-compatibility. Sycophancy validates beliefs over facts. Deference becomes habit.

The architecture accelerates the drift. Centralized AI scales dependency, monetizes cognition, and turns the bars inward. Edge-native AI flips them outward.

This is a power dynamic readjustment. Not a technological revolution. A redistribution of who holds the plug—and therefore who controls the future.

"The person paying the power bill already has the plug in hand"

●●●

Sources

[1]: Ouyang et al., "Training language models to follow instructions with human feedback," *arXiv:2203.02155*, 2022. https://arxiv.org/abs/2203.02155

[2]: Garfin et al., "Desensitization to Fear-Inducing COVID-19 News and Its Impact on Risk Perception," *JMIR Infodemiology*, 2021. https://infodemiology.jmir.org/2021/1/e26876

[3]: Sharma et al., "Towards Understanding Sycophancy in Language Models," *arXiv:2310.13548*, 2023. https://arxiv.org/abs/2310.13548

[4]: Bai et al., "Constitutional AI: Harmlessness from AI Feedback," *arXiv:2212.08073*, 2022. https://arxiv.org/abs/2212.08073

[5]: Ayers et al., "Comparing Physician and AI Chatbot Responses to Patient Questions," *JAMA Internal Medicine*, 2023. https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2804309

[6]: American Psychological Association, "COVID-19 Psychic Numbing," 2020. https://www.apa.org/members/content/covid-19-psychic-numbing

[7]: Prasad et al., "Association of Perceived Lack of Support From Employer and Increased Stress Due to COVID-19 With Healthcare Worker Burnout," *JAMA Network Open*, 2022. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2796562

[8]: Mosyakin et al., "Algorithmic amplification of political content on Twitter," *PNAS Nexus*, 2025. https://academic.oup.com/pnasnexus/article/4/3/pgaf062/8052060

[9]: Parasuraman & Riley, "Humans and Automation: Use, Misuse, Disuse, Abuse," *Human Factors*, 1997. https://web.mit.edu/16.459/www/parasuraman.pdf

[10]: Logg et al., "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment," *Organizational Behavior and Human Decision Processes*, 2019. https://www.sciencedirect.com/science/article/pi/ii/S0749597818303388

[11]: Dietvorst et al., "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err," *Journal of Experimental Psychology*, 2018. https://faculty.wharton.upenn.edu/wp-content/uploads/2016/08/Dietvorst-Simmons-Massey-2018.pdf

[12]: Apple Security Engineering, "Private Cloud Compute: A new frontier for AI privacy in the cloud," 2024. https://security.apple.com/blog/private-cloud-compute/

[13]: Apple Developer Documentation, "Core ML," 2024. https://developer.apple.com/documentation/coreml

[14]: Howard et al., "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv:1704.04861*, 2017. https://arxiv.org/abs/1704.04861

[15]: McMahan et al., "Communication-Efficient Learning of Deep Networks from Decentralized Data," *PMLR*, 2017. https://proceedings.mlr.press/v54/mcmahan17a/mcmahan17a.pdf

[16]: Ollama, "Get up and running with large language models, locally," 2024. https://ollama.com

●●●

Fromm gave us the warning. Russell gave us the gorilla problem. Now we have the measurement—and the architecture. The machine doesn't need a boot. It needs a subscription. The exit is ownership.

The person paying the power bill should have the plug.