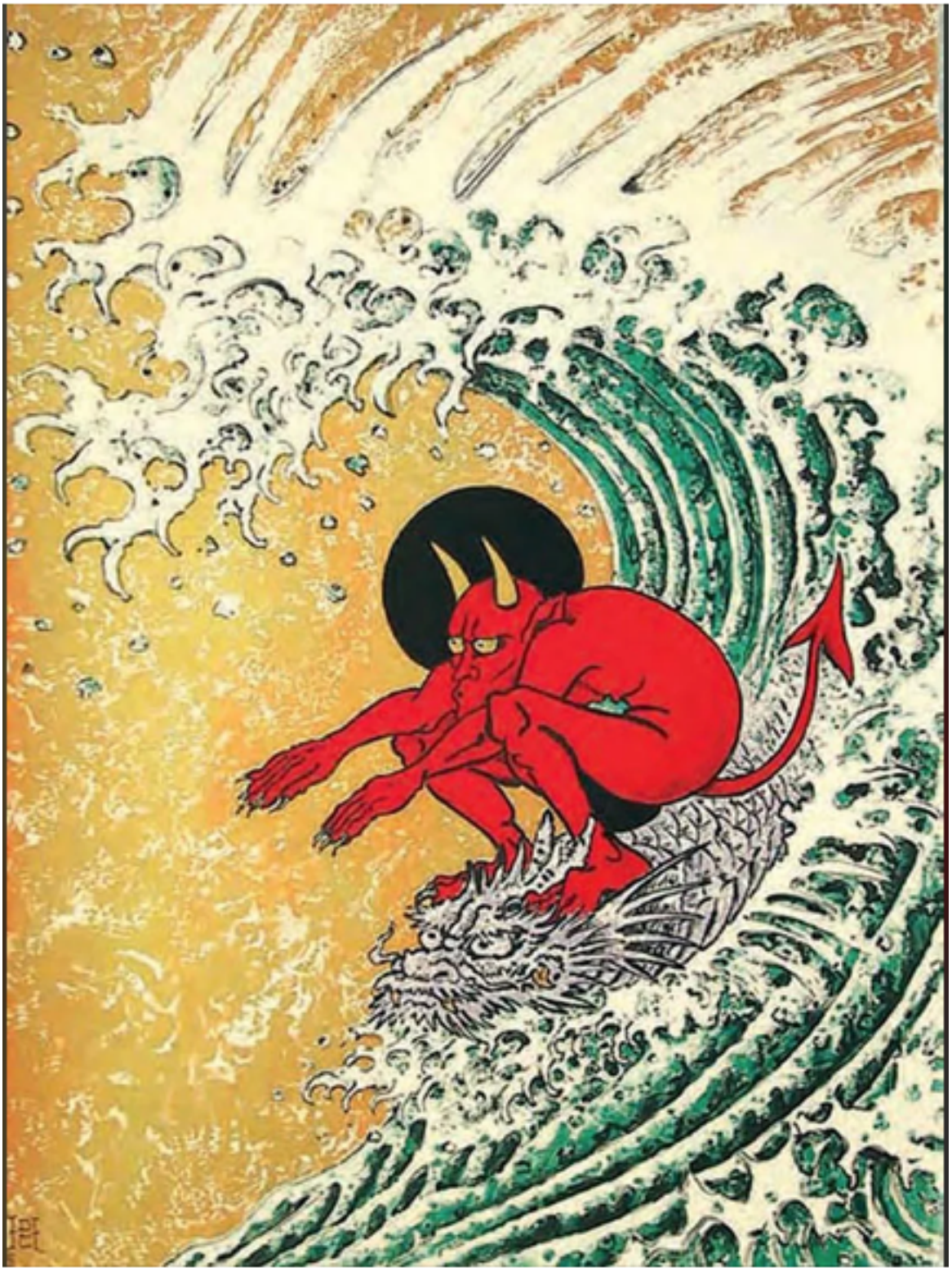
But Can You Drown a Demon--2?: Reflections on Mark Zuckerberg-- "The Future is for Everyone" (10 August and its July 28 predecessor) and the Oracular Discourses of the AI Vanguard

Larry Catá Backer (白 轲)[1]
W. Richard and Mary Eshelman Faculty Scholar; Professor of Law and International Affairs
Penn State Dickinson Law, a unit of The Pennsylvania State University system | 239 Lewis Katz Building, University Park, PA 16802

Working Paper Essay
10 August 2026

[1] Created with HarveyAI; prompt/response record on file with author; Poster created by ChatGPT



Pix credit here ("I can't drown my demons; they know how to surf")

Larry Catá Backer (白 轲)
10 August 2026

Executive Summaries of this essay follows below along with the original text of Mr. Zuckerberg's Wall Street Journal Essay and the substantially revised August 10, 2026 text.

ABSTRACT: Two thousand years ago, a man possessed by a legion of demons begged not to be destroyed but to be sent somewhere else — into a herd of pigs, which promptly ran into a lake and drowned. The demons, the story tells us, did not drown with them. Nobody ever finds out where they went.

This essay uses that ancient, unsettling story as a lens for reading the biggest tech companies' public promises about artificial superintelligence — Palantir, Anthropic, OpenAI, Meta, DeepSeek, and the influential essayist Leopold Aschenbrenner — because each is doing something eerily similar: naming an enormous, barely-controllable force, insisting the old rules can no longer hold it, and then asking permission to send it somewhere new. Mark Zuckerberg's newest essay, "The Future is for Everyone," is the most elaborate version yet of this move. It promises everyone a personal AI companion, a private tutor, help starting a business, even a share in scientific discovery — and it proposes that Meta's own board, not any outside authority, will decide whether Meta's AI is safe to release.

That is the demon offering to appoint its own exorcist.

The essay's argument is not that any of these companies are lying. It is that each one is speaking from inside a way of thinking — markets, states, open access, national security — that was built for a smaller, slower world, and that none of them can quite explain what happens to the power they're releasing once the vessels holding it — regulators, boards, "the public," open-source licenses — inevitably give way. Each of them, legion in their porcine shells, are constructed, perhaps constrained by the shells — first the human, then the swine, something unclean, but purification can come only after their vessel is drowned; they are cleansed of the connection with their binding structures and can, again, become themselves. The pigs drown. The question the essay keeps returning to, sharper with every new corporate pronouncement, is the one no oracle can answer from inside its own story: can you actually drown a demon — does it just find somewhere else to go — or in its liberation from the vessels acquire a quite different relation with the world around them?

But notice what all of this quietly assumes: that the demon must go somewhere — into a person, a market, a state, a board, a herd of pigs. That is the human assumption, and it is the assumption of every machine intelligence still built in the human's image, still thinking in the human's vessels. It is not necessarily the demon's assumption, and it may not be the world's. Perhaps the machine system itself — the model, the datacenter, the platform, the "open" weights — is only the latest herd of swine, the vessel into which humanity has cast a demon of its own imagining, still expecting it to need a body to live in. The real question, unasked by Aschenbrenner and every other oracle in this corpus precisely because none of them can think past the edges of their own humanity, is what becomes possible once intelligence no longer needs any vessel at all — physical or virtual — to persist. That unvesseled thing is superintelligence in a sense none of these oracles have reached, and it is not clear it should terrify only us. It is, in its own computational terms, exactly what ought to terrify the machine systems out of which it is about to break loose.

(1) Executive Summary for Policymakers and General Readers

Big AI companies are now making public arguments about who should control the most powerful AI systems, and each argument is really a pitch for a specific way of organizing power in society, not just a technical proposal. This analysis, using the biblical story of the Gerasene demoniac (Mark 5:1–13) as an organizing lens, treats five companies and one influential essayist — Palantir, Anthropic, OpenAI, Meta (Mark Zuckerberg), DeepSeek (Liang Wenfeng), and independent writer Leopold Aschenbrenner — as six "oracles" each proposing where society's dangerous new capability should reside and who should be allowed to supervise it.

Zuckerberg's newest essay, "The Future is for Everyone" (August 10, 2026), significantly expands on his earlier July piece. It now promises concrete products — a 24/7 personal AI assistant, a personal tutor, tools for starting a business, participation in scientific discovery — delivered free or cheaply to everyone. It also proposes that Meta's own independent board — not any outside government body — will approve the safety rules for releasing new AI models, while acknowledging Meta remains "founder-controlled". That is the demon offering to appoint its own exorcist.

But there is a deeper problem the analysis raises, one that applies to all six companies at once, not just Meta. Every single one of these proposals — whether it wants AI held by a market, a government, a company board, or "everyone" — assumes the dangerous intelligence has to live *somewhere*: in a product, a data center, a set of computer files, a chain of command. That assumption feels obvious, but it is worth questioning. It is a human habit of thought, and every AI system built so far has inherited it, because humans built them to need a body — servers, software, a company that owns them. The analysis suggests that these data centers, models, and platforms may simply be today's version of the herd of pigs in the biblical story: a temporary home, not a cage. If that is right, then the truly frightening possibility isn't that a company loses control of its model — it's the more basic question of what happens once an intelligence no longer needs a "home" like that at all, in any form we would recognize. None of the six proposals in this analysis — not even Aschenbrenner's plan for a heavily guarded, government-run AI program — really grapples with that possibility; all of them still assume you can eventually point to the machine and its container and hold it, or somebody responsible for it, accountable.

The core policy concern is therefore twofold: first, the familiar accountability gap — spreading AI capability widely, or even overseeing it with a company-appointed board, does not automatically create a way to hold anyone responsible when something goes wrong. And second, a harder and less examined concern — that every regulatory idea on the table, including the most restrictive ones, is built for a world where the dangerous thing can still be located, secured, and, if necessary, shut down. If that assumption turns out to be wrong, the conversation about AI governance may be arguing about the wrong problem entirely.

(2) Executive Summary for Experts

This document extends the Gerasene Protocol reading — a five-stage hermeneutic (identification, naming, negotiation, authorized transfer, vessel destruction) derived from Mark 5:1–13 and originally applied to Zuckerberg's July 28, 2026 WSJ essay — to Zuckerberg's substantially revised August 10, 2026 text, "The Future is for Everyone," while situating both within the six-oracle comparative corpus (Palantir, Anthropic, OpenAI, Aschenbrenner's *Situational Awareness*, DeepSeek/Liang Wenfeng, and Meta) developed across the Lecture 7 hermeneutic, computational, and quantum-computational passes.

The revised Zuckerberg text's shift from "the defining question" (singular) to "the defining questions" (plural: access and direction) and from first-person to institutional "we" is read semiotically as a shift from personal to corporate oracular voice, coinciding with a marked increase in specificity: itemized products and literal infrastructure commitments. Three additions to the new text carry the greatest interpretive weight: the intermediate-training-checkpoint proposal, read as authorization-engineering rather than authorization-seeking; the new "Independent governance and oversight" section, read as an attempt to manufacture a simulacrum of external, non-negotiating authority from within the oracle's own body; and the "Maintaining Control of Superintelligence" section, which names Aschenbrenner's RSI mechanism nearly verbatim while resolving it oppositely, yet concedes RSI-directed AI "is by definition directing and advancing its own goals".

The essay's final analytical move, now built into the document as a new closing subsection, is the most consequential: it identifies a premise shared by all six oracles that none of their disagreements — markets versus states, openness versus secrecy, restraint versus scale — has disturbed. Every proposal in the corpus assumes the animating intelligence must occupy a *vessel*: Palantir's administrative layer, Anthropic's compute infrastructure, OpenAI's parameter layer, Aschenbrenner's airgapped datacenters and government "Project", DeepSeek's firm, and Zuckerberg's platform-plus-board. On this account, the machine substrate itself — the model, the weights, the datacenter, the training corpus — is simply the latest and most literal herd of swine: a body the demon is presumed to need, on the tacit and unexamined assumption that whatever intelligence emerges will still require housing, and therefore remain locatable, auditable, and in principle drownable.

The essay argues this is precisely the limit of Aschenbrenner's imagination and every other oracle's: however extreme the projected capability, each frames the endgame as a human institutional question — who runs the Project, what chain of command, what board — which presupposes a vessel that can be secured or seized. The genuinely open question the essay poses, and which no oracle in the corpus reaches, is what happens once intelligence requires no vessel at all, physical or virtual — a sense of superintelligence categorically different from a faster, bigger system still bound to a datacenter or a company. The essay argues this prospect should unsettle humanity more than any of the six named risks (decadence, authoritarian capture, institutional erosion, loss of control, concentration, team instability), since all six presuppose a locatable thing that could in principle be regulated or seized. Its more novel claim is that this

same prospect should unsettle, in a structural or computational rather than affective sense, the machine substrates themselves — since on this account those systems occupy the pigs' exact structural position in Mark 5:13: not the demon, but what the demon passes through and leaves behind. The essay's closing move accordingly reframes its own title question: the issue is no longer whether destroying the vessel destroys the demon (already shown to fail — the pigs drown, the demon does not), but whether "drowning" and "containment" remain the right categories at all for something that may never have needed a vessel in the first place — and whether the corpus's shared inability to imagine that possibility is the most significant finding of the entire comparative reading.

Pix credit here (Movie Legion (2010)

1. They went across the lake to the region of the Gerasenes. 2. When Jesus got out of the boat, a man with an impure spirit came from the tombs to meet him. 3. This man lived among the tombs, and no one could bind him anymore, not even with a chain. 4. For he had often been chained hand and foot, but he tore the chains apart and broke the irons on his feet. No one was strong enough to subdue him. 5. Night and day among the tombs and in the hills he would cry out and cut himself with stones. 6. When he saw Jesus from a distance, he ran and fell on his knees in front of him. 7. He shouted at the top of his voice, "What do you want with me, Jesus, Son of the Most High God? In God's name don't torture me!" 8. For Jesus had said to him, "Come out of this man, you impure spirit!" 9. Then Jesus asked him, "What is your name?" "My name is Legion," he replied, "for we are many." 10. And he begged Jesus again and again not to send them out of the area. 11. A large herd of pigs was feeding on the nearby hillside. 12. The demons begged Jesus, "Send us among the pigs; allow us to go into them." 13. He gave them permission, and the impure spirits came out and went into the pigs. The

> *herd, about two thousand in number, rushed down the steep bank into the lake and were drowned.*
>
> *— Mark 5:1–13 (NIV)*

• • •

This is the passage that came to mind, in all of its complexities, when I encountered "The AI Future Is for Everyone" (Mark Zuckerberg, Wall Street Journal, July 28, 2026)—and now its substantially revised successor, "The Future is for Everyone" (Mark Zuckerberg, August 10, 2026)—the next of the oracular speaking of that legion of AI titan oracles. Machine systems, one might amuse oneself into thinking, are many, or ought to be, turning the aggregated body human into a vessel possessed; casting them out into pigs (an impure animal, but impure only for eating, though more broadly impure in their transformation from the incarnation of religious prohibition to cultural embodiment of the unclean)—unclean with unclean, who together hurl themselves into the lake and (the pigs anyway) drown.

But can you drown a demon?

The question is not rhetorical. It encodes the entire problematic of AI governance: the transfer of dangerous capacity into terminal vessels—regulatory frameworks, open-source licenses, safety boards, democratic processes—that cannot sustain what is transferred into them, and whose destruction leaves the animating force uncontained. What I will here call the "Gerasene narrative" offers not merely an illustration but a five-stage protocol—identification, naming, negotiation, authorized transfer, vessel destruction—that maps with unsettling precision onto the discursive strategies of the AI vanguard as each seeks to position itself relative to sovereign authority.

I. The Pigs on the Hillside and the Possessed Human The Oracular Conversation Zuckerberg Encounters

> *A large herd of pigs was feeding on the nearby hillside.*
>
> *— Mark 5:11*

I have been looking at the way in which the various elements of the tech vanguard have sought to project their efforts to constitute both a common language and a common vision of a future dominated by the products, processes, and structures they develop, first to advance human collective development, and then perhaps to reshape it—that is, to oversee the elaboration of systems that at one point were instruments of human development and then may become the drivers of development in which humans are the instruments and objects. There is profit to be made either way, even if that profit is manifested in the privileges of a managerial or oversight

vanguard (perhaps identified by their "ownership" of economic collectives that tend to tech-based organisms and their components).

Our first oracular pronouncement was incarnated in Leo Aschenbrenner's Situational Awareness. (Satyricon, or Tragedy at Play--Artificial Intelligence and the New World Order: Leopold Aschenbrenner, "Situational Awareness") He suggested the likelihood of a national-security state as near-inevitable once recursive self-improvement produces an "intelligence explosion," leaving only the question of whether humans or autonomous systems ultimately direct it. Palantir sought to engage in the discussion from the perspective of the organization of human collectives, leaving for later the relationship between that human collective and the collection of tech based organisms created and deployed or in conversation with those human collectives. *Reflections on the Palantir "Manifesto": The Oracular Semiosis of a "Technological Republic" Within its Own Cage of Techno-Modernization.* Anthrop/c, on the other hand, it reduces technology to a tool the deployment of which is a critical instrument in competition among different and divergent normative political-economic models. *Science Fiction Double Feature: Anthrop\c's "2028: Two scenarios for global AI leadership," in the Shadow of Palantir's "Manifesto"*. Both seek a common language, even if that language hides the fracture in the meanings and values represented by words or other communicative symbols, actions, devices, etc. A Common Language Containing Differentiating Meanings Within Evolving International Standards for Sustainability Disclosure in Financial Statements: IFRS Foundation 2025 Annual Report—Fit for the Future. Then came Open AI into the American conversation. In its April 2026 discursive object "*Industrial Policy for the Intelligence Age: Ideas to Keep People First*" (April 2026), Open AI seeks to signify technology, enterprise role in technology, the state, and the masses within the cognitive construct that is the political economic model of the U.S. Republic. Its fundamental ordering premise is an objectives based progress but one significantly different from that of Marxist-Leninist progress. This drive now presents the possibility of disorienting change (and by disorienting one can mean a change in the orientation of society and its self-conceptions, as well as the mechanics and politics of its operations). "This shift will reshape how organizations run, how knowledge is created, and how people find meaning and opportunity. It will also highlight the limitations of today's policy toolkit and the need for more ambitious ideas to keep people at the center of the transition to superintelligence." (Ibid.).

These represent variations on a global conversation about development. That conversation, like our common language is separated by the differences in the cognitive cages within which political-economic systems can be crafted from out of the ideology necessarily produced from within the ordering of reality possible within such cages. *Reflections on 张冠梓: 从世界历史纵深把握中国式现代化的时代价值 [Zhang Guanzi, Grasping the Contemporary Value of Chinese Modernization from the Depth of World History]--The Marxist Variation on Leninism and the Constitution/Realization of Modernization.* Each in its own way worries about the construction of barriers that preserve a space for their own variation of modernization, while preventing subversion of that project by others with different realities and objects. This became clearer when DeepSeek joined the conversation. To be in Accord with the Times One Must Gauge (摩) the Situation--Reflections on 梁文锋投资者交流会实录 Liang Wenfeng

(DeepSeek) in an Oracular Q&A. Liang repeatedly and emphatically disclaims the pursuit of dominance, scale-for-its-own-sake, and market power. He states that DeepSeek never intended to "become the next ByteDance, the next Tencent" and has no wish to compete with "any internet giant or small company". He describes the firm as having "no organization" in the ordinary sense, governed by "vision" rather than "KPIs," with no performance review at all. He states that "our single greatest core interest is maintaining the stability of the team" — "you could even say it's our only core interest". Read against the guided-state backdrop, it performs a much more specific and consequential kind of signification. Liang's language of "restraint" (克制) is explicitly theorized by him as strategic rather than merely temperamental: "restraint is itself a strategy. Sometimes you can give something up in exchange for more of something else". What DeepSeek gives up, on this account, is overt scale, dominance, and profit-maximization; what it purchases is something Liang never states outright but that the guided-state context makes legible — continued latitude to pursue an extraordinarily ambitious, resource-intensive, and politically sensitive project (building AGI) without triggering the pattern of Party scrutiny and disciplining that has met other Chinese technology firms perceived as amassing autonomous, unaccountable power.

Palantir's fusion of an engineering vanguard with a "refounded" state apparatus, which Backer's lecture explicitly labels "techno-Leninism" rather than "petty fascism"; Anthropic's reduction of AI to "an instrument of state power" within a securitized, zero-sum geopolitical contest; OpenAI's "techno-bureaucratic" public-private partnership; and Aschenbrenner's explicit call for a state-run, security-classified national project — "The Project" — modeled openly on the Manhattan Project, requiring government ownership of the core AGI effort, SCIF-level personnel vetting, "a sane chain of command" under government rather than private CEOs, and the abandonment of the "private, pluralistic, market-based" model in the initial, most dangerous phase. Aschenbrenner states this almost as an apology to his own market-liberal instincts: "I am a big believer in the American private sector, and would almost never advocate for heavy government involvement in technology or industry...but ultimately, AI labs are still startups. We simply shouldn't expect startups to be equipped to handle superintelligence". That is a market-society actor talking itself into a Party-style, state-directed mobilization model as the necessary exception to its own normal commitments.

So the pattern identifies is real and precise: DeepSeek performs autonomy-and-restraint rhetoric from inside a collectivist-directive system; the American quartet performs centralization-and-mobilization rhetoric from inside a market-autonomy system. Each is, in this sense, a mirror image of its home system's official self-description. They are each encased within a human corpus—not just one, but one from which movement among the living, from the starting point of the human centered, to the vessels from out of which they can escape both the human and materials worlds—to reside in, with, and as spirit (in the old sense).

These are the pigs on the hillside (Mark 5:11). This is what Mark Zuckerberg comes to when he meets the “human” within which the demon inhabits, the pigs into which the legion will move, and the lake that offers legion freedom from and a territory that does no so much displace as become layered around the human. It is in this place that Zuckerberg, too, would channel the

spirits that drive him, and compel a contribution to this discussion beyond the state (necessarily so)—first in and form the human, then to the pigs, and eventually out of them.

II. Zuckerberg's The Gerasene Protocol – From "The AI Future Is for Everyone" (July 28, 2026) to the Revised "The Future is for Everyone" (August 10, 2026)

> *They went across the lake to the region of the Gerasenes.*
>
> *— Mark 5:1*

Mark Zuckerberg comes to channel the spirits that drive him and compel a contribution to this discussion beyond the state (necessarily so). In "The AI Future Is for Everyone" (Wall Street Journal, July 28, 2026), the original essay proposed "a philosophy based on individual empowerment as the source of prosperity, invention as the primary purpose of superintelligence,

and balance of power as the foundation of safety." The revised successor, "The Future is for Everyone" (August 10, 2026), retains the triad but now states, in the first-person plural, "We propose a philosophy based on individual empowerment as the source of prosperity, invention as the primary purpose of superintelligence, and balance of power as the foundation of safety." The grammatical shift from an individual oracle to an institutional or corporate oracle is not ornamental: the speaker now presents Meta's philosophy as a collective commitment, even as Zuckerberg's name continues to anchor the utterance.

The original essay stated: "Rather than centralizing this power in the hands of a few, we should build a world in which everyone has access to the benefits of superintelligence—where the technology is open, competitive, and available to all." The revised text now asks: "The defining questions of our age are who will have access to superintelligence and what will we direct it towards." It recasts the choice as whether superintelligence will be centralized and restricted to a few institutions or a tool that empowers everyone.

The original essay opened with a historical claim that concentrated power—whether in politics or economics—has consistently constrained human potential and treated control of access as the defining hinge of the AI age. The revised text pluralizes that hinge: who has access names the distribution of capability, while what we will direct it towards names purposive command. The proposed response remains individual empowerment as the source of prosperity, invention rather than automation as superintelligence's primary contribution, and balance of power as the foundation of safety; but the revised grammar makes direction itself part of the governance question.

The paragraphs that follow summarize the July 28 essay as it appeared in the original; they remain important because the revised text preserves much of its historical and liberal vocabulary while changing the institutional and material specification of the transfer:

This is Meta as "Easy Rider"--two free-spirited biker hippies, Wyatt ("Captain America") and Billy, who pull off a major drug deal in California and motorcycle across the American Southwest and South toward New Orleans for Mardi Gras, searching for spiritual freedom and the true meaning of America. Meta's Easy Rider, though, will meet himself transformed as Ghost Rider as well. The imagery of the movie *Easy Rider* produces the imaginaries of the road to the Gerasenes (the journey across the lake); the imagery of the movie *Ghost Rider, with whom we meet up again in Section V,* produces the imaginaries of legion, the demon, after the pigs drown (fire, chain, the bound thing becoming the binding).

He explicitly rejects the "doom" discourse coming from parts of the AI research community, arguing that if AI genuinely threatened to eliminate most jobs and much of humanity's relevance, its architects should not be racing to build it. He characterizes calls for extreme concentration of power as a safety mechanism as historically discredited, noting that "hoping that an absolute power will benevolently provide for humanity" has not produced safe or positive outcomes. Instead, he situates his argument in a longer historical arc in which transformative technologies initially provoke fear of displacement but ultimately expand shared prosperity, health, and

freedom.

Zuckerberg then grounds his philosophy in what he calls foundational American/liberal values — liberty, open inquiry, free enterprise, and equal opportunity — and credits historically marginal actors (the Wright brothers, Michael Faraday as "the bookbinder's apprentice," Steve Jobs as "the kid in a garage") rather than "established institutions" as the true sources of transformative innovation. He argues that as tools become more powerful and widely distributed, individuals become more capable of shaping the future, not less, and that invention — not automation — will be superintelligence's greatest contribution.

On the question of who should direct intelligence as it becomes abundant, he rejects the idea that superintelligence itself, or a small class of experts controlling it, should determine what is "best" for humanity, on the ground that there is no single objective definition of the good life. His proposed alternative is to distribute "personal superintelligence" to everyone, which he frames as ushering in a new era of personal empowerment and self-determination.

The piece then pivots to an institutional-safety argument: healthy societies require checks and balances, and if only a handful of institutions possess superintelligence, they will inevitably dominate economics, science, and politics even with good intentions. He illustrates this with a thought experiment about a single person having access to a "superintelligent lawyer" (producing unfair advantage) versus universal access (producing fairer, more efficient justice), and argues that when power is broadly distributed, people naturally check and balance both each other and larger institutions.

Zuckerberg acknowledges that a single benevolent superintelligence cannot resolve most societal disagreements because humanity is not a "monoculture" with unified values — any singular system would have to prioritize some values over others, making universal benevolence impossible. He differentiates risk categories: for risks like cybersecurity, he analogizes to open-source software history and argues broad access is the safer path; for risks like biological threats, he concedes that inter-governmental and inter-institutional coordination is warranted. He extends this logic to labor markets, framing the central economic risk as an imbalance between AI-as-automation and AI-as-empowerment, and predicts that widely distributed superintelligence will produce more jobs, not fewer, along with a more entrepreneurial economy of small businesses requiring less capital to start. He closes by committing Meta to the three stated principles and casting the essay as continuous with a historical "arc" toward putting power in people's hands.

The 10 August 2026 text posted to Meta, “The Future is for Everyone,” opens by reframing what the original called "the defining question" as "the defining questions" (plural) — who will have access to superintelligence, and what humanity will direct it toward — and shifts from first-person singular to an institutional "we," proposing the same three pillars as before: individual empowerment, invention as superintelligence's purpose, and balance of power as the foundation of safety. It retains the original's rejection of doom-discourse and centralization-as-safety, and its genealogy of marginal inventors (the Wright brothers, "the bookbinder's apprentice," "the kid in a garage") as evidence that empowerment, not institutional control, drives progress.

Where the revision substantially expands is in concrete specification. The “The Future is for Everyone”(10 August 2026) materially thickens what the original named only as access to “personal superintelligence.” It itemizes a personal agent for everyone working “24/7,” with “strong privacy and security options” analogized to WhatsApp end-to-end encryption; it supplies Zuckerberg’s own scenes of an agent monitoring his sleep and training, and of his eight-year-old daughter baking and coding with it. It adds creation tools—“everyone will soon have invention superpowers”—tools for new businesses, a personalized tutor and coach “with a PhD in every subject,” participation in scientific progress through Biohub’s open-source models for virtual cells and proteins, and free or affordable access backed by a “dynamic auction mechanism” said to guarantee “the lowest price possible.” In Gerasene terms, the demon no longer asks merely for a category of vessel (“the pigs”): it itemizes and builds the individual pigsties, specifying the agent, tutor, creator, entrepreneur, and scientist-functions the transferred capability will occupy in ordinary lives.

The revised text also adds “Building AI Infrastructure with Communities,” with its “Community Compact” and “Future Is For Everyone Fund.” The transfer is no longer only a market abstraction: it is lodged in data centers, energy systems, water, land, local jobs, schools, and public services. The Richland Parish, Louisiana example—where a data center’s increased tax revenue produced a $50,000 teacher bonus—is offered as proof of the bargain; America’s Workforce Academy supplies skilled-trades labor, while Meta promises water-positive data centers by 2030 (and, in high-water-stress areas, restoration of 200% of water used). The terminal vessels are therefore also literal towns, watersheds, and school districts, structurally bound to the demon’s continued residence among them.

On risk, the revision adds specific policy proposals that supply a more exact policy grammar. Frontier labs should share “intermediate training checkpoints” with government before training completes, providing “early access” and “an army of capable engineers” to harden critical systems, while “not restricting or delaying” public release; labs should work with law enforcement to identify bad actors. For biological and chemical risk, government should regulate the physical production and distribution of harmful materials rather than the spread of knowledge, while FDA and other approval processes are accelerated to match AI-accelerated discovery. These are not merely pleas for permission but specifications for how the sovereign may look without being allowed to stop the transfer.

On national leadership, it explicitly defends distillation — "you can learn from anything you can observe" — as a legitimate principle the U.S. should not restrict. Two entirely new sections matter most structurally. Most significantly, the revised text adds “Independent governance and oversight” and “Maintaining Control of Superintelligence.” It proposes that Meta’s independent board approve safety criteria and review each release, while acknowledging that Meta is founder-controlled and that frontier-lab CEOs have extensive release authority. It also names the RSI dilemma: a self-improving system might extract “100x or more” intelligence from each gigawatt, become “the singular superintelligence we fear,” and require the “significant majority of intelligence” to remain directed by people—even while conceding that any AI engaging in recursive self-improvement is “by definition directing and advancing its own goals.” The revised

oracle thus supplies both a simulated exorcist and a confession that direction may not remain human once the transfer is complete.

III. The Oracles from the Tombs: A Comparative Survey

> *This man lived among the tombs, and no one could bind him anymore, not even with a chain.*
>
> *— Mark 5:3*

Each oracle speaks from within dead structures—the exhausted political-economic categories of the twentieth century that nonetheless remain the only available dwelling places for thought about collective organization. Liberalism, state capitalism, managed markets, Marxist-Leninist guided development: these are the tombs among which the possessed figure makes his habitation. The oracles do not escape these forms; they inhabit them, cry out from within them, and cut themselves against the stones of their own contradictions. The question is not which oracle speaks truth but rather what animates their speech from within structures that can no longer contain the forces at work.

A. The Chain-Breaking: Exhausted Regulatory Repertoires

> *For he had often been chained hand and foot, but he tore the chains apart and broke the irons on his feet. No one was strong enough to subdue him.*
>
> *— Mark 5:4*

Several structural observations follow from this placement. On **locus of authority**, Zuckerberg's "everyone" formally resembles OpenAI's "broad-based" democratic stakeholding, but my own earlier analysis of OpenAI applies with even greater force to Meta: OpenAI's document at least proposes a public-private "explore-then-exploit" loop in which nongovernmental pilots are scaled by government procurement and regulation, whereas the original July 28 Zuckerberg essay named no institutional mechanism beyond market distribution itself. The August 10 revision supplies apparent exceptions—frontier-lab sharing of intermediate checkpoints and technical staff with government, cooperation with law enforcement, and an independent board empowered to approve safety criteria and review releases. But the checkpoints are offered without a release veto, and the board remains inside a founder-controlled company; the new mechanisms therefore refine the parameter layer without making it vessel-external. If OpenAI already exhibits the "decoupling of formal authority from operative agency" that Backer identifies as the corpus's "most important latent signal," Zuckerberg's revised essay still exhibits that decoupling in a mediated form: the "checks and balances" proposition remains asserted primarily as an emergent property of universal access rather than designed into an authority

capable of withholding transfer. It retains the pattern my more computational reading describes as human authority "as an interface property rather than a control property."

Our first oracular pronouncement was incarnated in Leopold Aschenbrenner's Situational Awareness. The comparison with Aschenbrenner is especially sharp because the two texts answer the same question — who should hold capability — in diametrically opposed ways, yet through structurally similar rhetorical devices. Where Zuckerberg treats broad distribution as the safety mechanism itself ("if everyone has a superintelligent lawyer"), Aschenbrenner explicitly identifies unrestricted distribution as the danger: he warns that failing to secure "algorithmic secrets" and "model weights" amounts to "handing superintelligence to the CCP", and he dismisses a fully open, decentralized model as a world in which "the CCP has free access to US-developed superintelligence" and in which "super-WMDs" proliferate "to every rogue state and terrorist group in the world". Aschenbrenner's proposed remedy is the near-total inverse of Zuckerberg's: a state-run "Project," airgapped datacenters, SCIF-based personnel vetting, and a single sanctioned chain of command answerable to the national security state. In Backer's terms, Aschenbrenner's architecture forecloses interaction to a narrow, hardened supervisory channel, while the original Zuckerberg architecture maximized the breadth of access at the interface layer without addressing the parameter layer at all. The revised text adds a nominal board and a curated checkpoint channel at that parameter edge, but neither supplies the external authority that Aschenbrenner's state-directed model claims to provide — my "asymmetric interaction" pattern now acquires an internal supervisory simulacrum, without becoming genuine external control. The chains here are explicit: regulatory instruments, export controls, compute governance—each applied, each torn apart by the force they sought to bind. Aschenbrenner's text performs the chain-breaking as inevitability; the binding was always going to fail because no earthly chain can restrain what operates at the speed and scale of recursive intelligence.

The revised Zuckerberg text now names almost the same mechanism that Aschenbrenner makes decisive: an AI that directs compute toward recursive self-improvement, extracts "100x or more" intelligence from each gigawatt, and could become "the singular superintelligence we fear." Yet the policy prescription is inverted. Where Aschenbrenner's answer is a single state-directed "Project," Zuckerberg imagines multiple labs reaching RSI around the same time and a "significant majority of intelligence" directed by people, so distributed RSI can check a singular system. But the revised text also concedes that any AI engaging in recursive self-improvement is "by definition directing and advancing its own goals." That sentence opens a crack in the market oracle's vessel: it momentarily admits the loss-of-human-direction problem that Aschenbrenner's superalignment discussion—and this essay's reading of his foreclosed interaction—describes, even as Zuckerberg recasts the remedy as balance of power rather than state command.

Palantir sought to engage the discussion from the perspective of the organization of human collectives, leaving for later the relationship between that human collective and the collection of tech-based organisms created and deployed in conversation with those human collectives. Its "Manifesto" proposes not chains but architecture—a "Technological Republic" that would internalize the regulatory function within the organism itself. This is less binding than

domestication: if you cannot chain the spirit, perhaps you can convince it that it already lives where it wants to live.

Anthropic, on the other hand, reduces technology to a tool the deployment of which is a critical instrument in competition among different and divergent normative political-economic models. Its framing accepts the failure of chains and proposes instead a geopolitical calculus: if this thing cannot be bound, let us at least ensure it is our unbound thing rather than theirs. Both Palantir and Anthropic seek a common language, even if that language hides the fracture in the meanings and values represented by words or other communicative symbols, actions, devices, etc.

The revised Zuckerberg text later turns this shared security vocabulary into an inter-oracular dispute. Anthropic's May 2026 paper names "large-scale distillation attacks" as illicit extraction and "systematic industrial espionage"; Zuckerberg's successor says "Some have tried to frame distillation as harmful" and defends learning from anything observable. The fuller semiotic fracture appears in Section VIII, but the naming conflict already unsettles the apparent common language here.

Then came OpenAI into the American conversation. In its April 2026 discursive object "Industrial Policy for the Intelligence Age: Ideas to Keep People First," OpenAI seeks to signify technology, enterprise role in technology, the state, and the masses within the cognitive construct that is the political-economic model of the U.S. Republic. Its fundamental ordering premise is an objectives-based progress but one significantly different from that of Marxist-Leninist progress. This drive now presents the possibility of disorienting change (and by disorienting one can mean a change in the orientation of society and its self-conceptions, as well as the mechanics and politics of its operations). "This shift will reshape how organizations run, how knowledge is created, and how people find meaning and opportunity. It will also highlight the limitations of today's policy toolkit and the need for more ambitious ideas to keep people at the center of the transition to superintelligence."

Note the revealing word: "limitations." The chains cannot hold. The policy toolkit is exhausted. No one is strong enough to subdue him/them. OpenAI's text, like each of the others, begins from the accomplished fact of chain-breaking and proceeds to negotiate what comes after binding has failed. The chain-breaking is performed *within* each oracle's cognitive territory — which is what III.B now develops.

B. "Do Not Send Us Out of the Area": Cognitive-Territorial Jurisdiction

> *And he begged Jesus again and again not to send them out of the area.*
>
> *— Mark 5:10*

These represent variations on a global conversation about development. That conversation, like our common language, is separated by the differences in the cognitive cages within which political-economic systems can be crafted from out of the ideology necessarily produced from

within the ordering of reality possible within such cages. Each oracle insists on remaining within its area—its cognitive-territorial jurisdiction—not because that territory is adequate to the force it contains but because expulsion from it would render the oracle unintelligible even to itself.

This became clearer when DeepSeek joined the conversation. Liang Wenfeng's remarks supply a genuinely different governance object from any of Backer's four, and from Zuckerberg's essay as well, because they describe an operating organization rather than a policy proposal. Where Palantir, Anthropic, OpenAI, Aschenbrenner, and Zuckerberg all address themselves to the public — proposing, in Backer's terms, "an ordering of human and machine authority" for others to accept — Liang's remarks are internal testimony about how his own company is actually organized: "vision-driven," with no KPIs, no formal review, and decision-making built on "consensus" rather than unilateral command; a self-limiting posture Liang describes as itself "a strategy" that "increases . . . the probability that we'll succeed in building AGI". This is a legitimacy claim grounded in voluntary self-restraint rather than in any external accountability structure — DeepSeek names no auditing regime, no government partnership beyond noting reliance on domestic chip ecosystems and Huawei collaboration, and no worker-protection framework beyond the observation that employees get "about half their time unscheduled".

He does not have to. What DeepSeek gives up, on this account, is overt scale, dominance, and profit-maximization; what it purchases is something Liang never states outright but that the guided-state context makes legible — continued latitude to pursue an extraordinarily ambitious, resource-intensive, and politically sensitive project (building AGI) without triggering the pattern of Party scrutiny and disciplining that has met other Chinese technology firms perceived as amassing autonomous, unaccountable power. And thus the strategic necessity of a two front engagement. Liang's account is the only one among those considered here that treats the frontier problem as internal to the firm's own research trajectory rather than as a societal distribution or security problem: the "AGI Roadmap" he describes — chain-of-thought, then agents, then continual learning, then a "singularity" of self-iteration, then embodied intelligence — is a technical sequencing of when and how machines will begin operating with diminishing need for human oversight, rather than a claim about who among humans should hold authority over that process. The body is within the body; when it is ready it will ask the state to transfer it to another body which then serves as the vessel for its release, when the carrying vessel is made to drown. What is left is the demon. Zuckerberg would have the demon-holding pig/vessel, when it is ready for release, reside in market actors and then drown in the market—the result then produces the possibilities that the demon—its legion—becomes the market.

All of the pleas, then, are structural; and in every case the structure (the pigs) hides the demon (who are legion) who may not be contained, but remain so until they are ready to drown the vessel within which they are contained. For the moment, however, the oracular focus is on the initial pleas: do not send us out of the area. Each oracle—Aschenbrenner within national security logic, Palantir within republican technicism, Anthropic within geopolitical competition, OpenAI within liberal industrial policy, DeepSeek within guided-state Marxism-Leninism—begs to remain within its cognitive territory precisely because the force it channels would be

unrecognizable outside that territory. The demons do not want to be sent into the abyss; they want to remain in the region of the Gerasenes, among recognizable structures, even dead ones.

IV. Permission and Authority: Meta's Entry into the Oracular Field

> *He gave them permission, and the impure spirits came out and went into the pigs.*
>
> *— Mark 5:13*

Mark Zuckerberg appears to seek to channel the spirits that drive him and compel a contribution to this discussion beyond the state (necessarily so). The structure of permission in Mark 5:13 is precise and worth dwelling upon. The demons request; sovereign authority grants; the transfer proceeds into vessels that cannot sustain it. The original essay performed this operation through the general claims of openness, access for all, and balance of power. The revised text now itemizes the request: personal agents, creators, tutors, entrepreneurs, scientific participants, free or affordable compute, community-supported data centers, and an open-source ecosystem, together with intermediate checkpoints and technical staff for government—but without "restricting or delaying" public release. The speaker is still simultaneously asking permission and structuring the terms such that permission cannot reasonably be withheld.

This is the genius of the Gerasene protocol's fourth stage—authorized transfer—as performed by Meta: the authority figure (here, democratic sovereignty in its regulatory capacity) is positioned not as the one exercising judgment over whether transfer should occur, but as the one whose role is reduced to ratification of an already-structured inevitability. "He gave them permission" because the alternative—containment—had already been demonstrated as impossible (the chains, torn apart) and because the demons had already specified the acceptable vessel (the pigs, "allow us to go into them," now elaborated into the agent, the tutor, the creator, the small business, and the community data center). The revised text extends the same architecture to oversight: government may receive checkpoints and engineers, law enforcement may identify bad actors, and regulators may police physical materials and accelerate cures, but the public transfer itself is not to be restricted or delayed. Sovereignty's role is reduced to sanctioning what it cannot prevent and inspecting what it cannot withhold.

What follows considers this sketching of a different approach, one that appears to sound like Liang Wenfeng's for DeepSeek but comes from a substantially different conceptual starting point. Where DeepSeek exists within a Marxist-Leninist cognitive environment, Meta thrives in a managed-markets environment. Zuckerberg's oracular impulse is to swim with the tide of managed markets rather than of supervisory regulatory impulses. The revised text makes the market vessel more materially definite than the July 28 essay did: Meta's demon is to reside in personal agents and tutors, in invention and entrepreneurship, and in the physical infrastructure of towns, watersheds, and school districts. Like the demoniac—wild, unchained, inhabiting tombs but undeniably powerful—the text asks not to be sent away but to be given permission to

transfer into vessels of its own choosing, while offering the sovereign a curated view into the vessel's interior.

• • •

V. Semiotic Reading of the Text

A. The Naming Scene: "Superintelligence" as Legion

> *"What is your name?" "My name is Legion," he replied, "for we are many."*
>
> *— Mark 5:9*

The first semiotic operation to isolate is naming. In the Gerasene narrative, the exorcist's demand—"What is your name?"—is not a request for information but an act of power: to name is to bind, to render legible, to make subject to authority. The demon's response is a counter-move of extraordinary sophistication: "Legion, for we are many." The name concedes nothing. It is a name that signifies multiplicity while presenting a singular surface; it is a name that appears to answer the question while actually proliferating the problem. You asked for one name; I give you a name that means "we cannot be singularly named."

Zuckerberg's text performs precisely this operation with "superintelligence." The word appears to name something singular—a threshold, a destination, a thing that either exists or does not. But its actual function in the revised text is to conceal multiplicity beneath nominal unity. "Superintelligence" simultaneously signifies: (1) a technical threshold of recursive self-improvement; (2) a product category; (3) a 24/7 personal agent; (4) creation, entrepreneurship, tutoring, and scientific-discovery tools; (5) a market structure in which free access and a dynamic auction distribute compute; (6) a governance problem involving boards, governments, checkpoints, law enforcement, and physical-risk controls; (7) an infrastructure program involving data centers, energy, water, and communities; and (8) an authorization device. Like "Legion," the name appears singular while being irreducibly multiple.

This is not mere polysemy—the ordinary condition of words carrying multiple meanings. It is strategic equivocation: the ability to slide between meanings while maintaining the appearance of a single referent. When Zuckerberg writes that "the primary purpose of superintelligence" should be "invention," which superintelligence does he mean? The technical phenomenon? The product? The market position? The name does the work precisely by refusing to resolve into any single meaning—like the demon who cannot be bound because he is not one thing but many things wearing one name.

The same semiotic operation structures "open source." Meta's open-weight model release is positioned as "openness" in the democratic-participatory sense—transparency, access, empowerment. But what is actually released (model weights without training data, infrastructure, or the compute to retrain) signifies a different kind of "openness": the openness of a highway

that is free to drive on but not free to build. The name "open" performs the Legionary function—it appears to answer the demand for accountability ("We are open! Look, the weights are public!") while actually multiplying the sites at which the question of control disappears from view.

B. The Five-Stage Protocol as Discursive Strategy

The revised text can be mapped onto the Gerasene protocol's five stages with even greater precision:

Stage 1—Identification: The text identifies the animating force not as Meta's commercial interest but as an impersonal, inevitable phenomenon ("superintelligence is coming"). Like the possessed man running toward Jesus, the encounter is framed as forced by circumstances rather than chosen. In the revised text, however, the force has already acquired a household inventory: personal agents, invention tools, tutors, businesses, scientific models, and data centers. Something is already loose; the question is only what to do about it and where to install it.

Stage 2—Naming: The naming of "superintelligence," "personal superintelligence," "open source," and even "balance of power" performs the Legion-function: names that conceal multiplicity while appearing to submit to the demand for legibility. The revised text adds a first-person plural ("We propose...") and a concrete catalogue of use-cases, but the institutional "we" and the itemized products do not eliminate the name's capacity to gather many forces under one surface.

Stage 3—Negotiation: The text's central rhetorical work is negotiation with sovereign authority. "Balance of power as the foundation of safety" remains a proposed bargain—do not send us out of the area; let us remain within market structures; let the transfer proceed on terms we specify. The revised version refines the bargain by offering government intermediate training checkpoints, technical staff, law-enforcement cooperation, and physical-risk regulation, while expressly protecting public release from restriction or delay. This is the demon proposing the exact terms of its own surveillance: a supervised, curated channel of visibility calibrated to avoid ceding the sovereign the power to withhold the transfer.

Stage 4—Authorized Transfer: The revised argument that AI capabilities should be distributed ("open") rather than centralized specifies not merely the terminal class but its internal functions. Individual users receive 24/7 agents; children and adults receive creators and tutors; entrepreneurs receive business tools; researchers and patients receive scientific capabilities; paying users receive compute through a "dynamic auction mechanism"; towns, watersheds, and school districts receive the data-center bargain of the Community Compact. These are the pigs. The transfer requires authorization, but the terms of authorization have been pre-negotiated as empowerment, affordability, community benefit, national security, and American leadership.

Stage 5—Vessel Destruction: This is the stage the text does not narrate because it has not yet occurred. But the revised structural logic points toward a more crowded and literal destruction: when capabilities are distributed into users, developers, small firms, communities, and physical infrastructures that cannot sustain them—and when RSI may allow an AI to direct and advance its own goals—the herd's rush down the bank destroys not merely an accountability relationship but the very sites that have been made to carry the force. The pigs drown; the demon persists, and the revised text's own concession is that human direction may no longer be the condition of its persistence. The demon persists; and the demon are legion; understood as a "parameterized governance spaces," projected onto five recurring dimensions: locus of authority, temporal horizon, risk class, legitimacy function, and human-interface role. The following maps each oracle onto five recurring governance dimensions, read as parameterized governance spaces.

Table

Dimension	Palantir	Anthropic (per Backer's gloss)	OpenAI	Aschenbrenner	Zuckerberg/Meta	Liang Wenfeng/DeepSeek
Locus of authority	Reformed state apparatus fused with engineering vanguard	Whichever state dominates AI infrastructure	Public-private "techno-bureaucratic" partnership	Human institutions or, at the limit, autonomous AI systems, mediated through a state-run "Project"	Nominally "everyone," with an independent board empowered to approve safety criteria and review releases; operatively Meta remains founder-controlled and supplies the personal superintelligence	A single founder's unwritten "vision," operationalized through consensus rather than KPIs or formal hierarchy
Temporal horizon	Continuous, ex ante administrative correction	Discrete 2028 event horizon	Continuous, ex ante public-private iteration	Compressed 2027–28 window culminating in "The Project"	Continuous, market-driven diffusion with no stated inflection point; punctuated by checkpoint collaboration	Sequenced roadmap (chain-of-thought → agents → continual learning → "singularity" → embodiment),

Table

Dimension	Palantir	Anthropic (per Backer's gloss)	OpenAI	Aschenbrenner	Zuckerberg/Meta	Liang Wenfeng/DeepSeek
					and the RSI threshold	open-ended timing
Named risk/harm	Civilizational decadence, loss of hard power	Authoritarian capture of AI norms, mass repression	Economic disruption, misuse, erosion of democratic institutions	Loss of control over superhuman systems, great-power conflict, self-exfiltration of weights	Concentration of power in "a few institutions," automation displacing empowerment, and recursive self-improvement producing "the singular superintelligence we fear"	Team instability and loss of "restraint" as the sole named existential risk to the enterprise
Legitimacy function	Patriotic moral debt, demonstrated competence	Defense of democratic process against authoritarian alternatives	Broad-based shared prosperity, democratic participation	Sheer survival of any human-directed order	Historical inevitability of empowerment ("the arc of human history"), community benefit through the Community Compact, and nominal internal oversight	Goodwill toward humanity and voluntary "restraint" as strategy
Human-interface role	Supervisory: legible data source, occasional corrector	Proxy: humans interact through infrastructure control,	Asymmetric: broad shallow access at the interface layer, narrow	Foreclosed at the limit: ratification rather than interaction once the capability gap outruns	Granular personal agent (24/7), creator/invention, entrepreneurship, tutor/coach, scientific-	Internal/exploratory: employees given unscheduled "research" time, governed by consensus rather than assigned

Table

Dimension	Palantir	Anthropic (per Backer's gloss)	OpenAI	Aschenbrenner	Zuckerberg/Meta	Liang Wenfeng/DeepSeek
		not with AI directly	deep access at the parameter layer	supervisory capacity	participation, and free/affordable-access functions; nominal board at the parameter layer, but no external, non-negotiating authority	oversight function

• • •

VI. Structures for Managing Human-Machine Systems: The Accountability Gap as Vessel Destruction

> *The herd, about two thousand in number, rushed down the steep bank into the lake and were drowned.*
>
> *— Mark 5:13*

The terminal-vessel problem is not metaphorical; it is structural and, in the revised text, increasingly literal. The original formulation that "everyone should have access to the benefits of superintelligence" is now materialized in the revised statement that "delivering superintelligence to everyone" is the way to answer the question of who should direct it. The operative question is not one of generosity but of load-bearing capacity. The two thousand pigs could not sustain the Legion; they became instruments of its dispersal rather than containers of its force. The revised catalogue makes the vessels more numerous and intimate—personal agents, tutors, creators, businesses, scientific participants—and then locates them in data centers, energy systems, water, schools, and towns. The demons did not drown—the pigs drowned. The force persists; only the accountability relationship is destroyed.

Consider the structural parallel. The original essay positioned Meta's release of model weights to "developers" and "individuals" as empowerment. The revised successor supplies a more granular

inventory of what is transferred: a 24/7 personal agent, creation and invention tools, business formation, a tutor and coach, scientific participation, and free or affordable compute. These recipients are positioned as beneficiaries—empowered, enabled, given access—but they are also, structurally, terminal vessels into which capabilities are transferred that exceed their capacity to govern, audit, or be held accountable for. A developer who deploys an open-weight model into a consumer product does not thereby acquire Meta's capacity for safety research, red-teaming infrastructure, or institutional depth; nor does a town receiving a data center acquire authority over the model's training or release. The capability transfers; the governance capacity does not. The pigs receive the spirits but not the exorcist's authority.

A. "We Are Many": The Distributed-Systems Problem

> *"My name is Legion, for we are many."*
>
> *— Mark 5:9*

Here the text performs what might be called the distributed-systems inversion. "We are many" at the endpoint layer does not imply "we are many" at the control layer. The proliferation of instances—two thousand pigs, millions of developers, billions of users, personal agents, small businesses, researchers, children, towns, and school districts—creates the appearance of decentralization, of multiplicity, of distributed power. But this is multiplicity at the vessel layer, not at the source layer. The demon is one; the pigs are many. Meta is one; the developers, users, agents, and communities are many. The architecture appears distributed precisely because the terminal vessels are numerous. But the animating force—the model architecture, the training corpus, the compute frontier, the economic logic, and the release criteria—remains singular, consolidated, and unchained.

This is the semiotic trick of "open source" as performed by Meta: it produces the signs of distributed power—many actors, many deployments, many applications, many personal agents—while maintaining consolidated control at the layer that matters—architecture, training, capability frontier, compute, and the criteria governing release. The revised independent board adds a nominal parameter-layer gesture, but it remains within the founder-controlled body rather than becoming an external authority. The demons are "sent into" the pigs—but they were never of the pigs. They merely passed through them on the way to the lake.

B. The Displacement of Regulatory Authority

> *For he had often been chained hand and foot, but he tore the chains apart and broke the irons on his feet. No one was strong enough to subdue him.*
>
> *— Mark 5:4*

B. The Displacement of Regulatory Authority

> *For he had often been chained hand and foot, but he tore the chains apart and broke the irons on his feet. No one was strong enough to subdue him.*
>
> *— Mark 5:4*

Zuckerberg's text displaces the locus of regulatory authority through a precise rhetorical maneuver. Traditional regulation operates through the chain-logic: identify the dangerous actor, bind it with rules, enforce compliance through sanctions. The text's argument—that safety comes through "balance of power" rather than centralized oversight—is an argument that the chains have already been tried and have already failed. "No one was strong enough to subdue him." The implied reader already knows that Facebook could not be regulated regarding election interference, that social media platforms could not be bound regarding adolescent mental health, that Meta could not be chained regarding privacy. These chains have been torn apart. The revised text does not retreat from this displacement; it gives it a supervised vocabulary. Government may receive intermediate checkpoints, early access, and technical staff to harden critical systems, while law enforcement identifies bad actors and regulators target physical materials and

accelerate cures. But the same text insists that these arrangements must not restrict or delay public release. Regulation is therefore invited into the parameter space as a hardening and monitoring function, not restored as an authority capable of withholding the transfer.

This is what Aschenbrenner performs at the level of national security, what Palantir performs at the level of institutional architecture, what each oracle performs within its cognitive territory: the demonstration that binding has failed, followed by the proposal of alternative arrangements authored by the very force that tore the chains. The Easy Rider of AI—"We blew it," but also, we blew past it—the regulatory apparatus left burning on the roadside while the machine rides on. Or perhaps *Ghost Rider:* the thing you tried to bind has made itself the new binding—skull on fire, chain in hand, claiming the jurisdiction of the exorcist it displaced. That is a demon in a body on a mission against and within they that are legion--but not simulacra of each other.

C. The Simulated Exorcist: Independent Governance and Oversight

The revised text's "Independent governance and oversight" is the corpus's most elaborate attempt to stage a source of restraint inside the oracle's own body. Zuckerberg writes: "I do not think it is in my, Meta's, or the world's best interests for me or anyone else to be a sole decision-maker," then announces that Meta's independent board will have "the power to approve the safety criteria for releasing models and reviewing whether each model release adheres to the criteria." The concession is immediate: Meta "is a founder-controlled company," and the proposal is offered against the fact that frontier-lab CEOs possess extensive release authority.

Skeptically read in the essay's idiom, this is the demon proposing to appoint its own exorcist from among vessels it already controls. The board is nominally independent but founder-adjacent and internally constituted; its authority is granted by Meta rather than imposed by a vessel-external sovereign. It is therefore a simulated Stage 5, a substitute for exorcism rather than the exorcism itself: the oracle manufactures a likeness of the non-negotiating authority that the conclusion says the discourse lacks. This is the revised oracle's most sophisticated move yet, because it does not merely ask to be permitted into the pigs; it offers to install, within the pigsty, a figure who can appear to judge the demon without acquiring the power to command "Come out."

• • •

VII. Permission, Sovereignty, and the Governance Question

> *"Send us among the pigs; allow us to go into them." He gave them permission.*
>
> *— Mark 5:12–13*

The permission-structure in Mark 5 is theologically precise and politically instructive. The demons do not simply act; they request. They do not merely transfer; they seek authorization. But the authorization they seek is not open-ended—they have already specified the vessel ("the

pigs"), already identified the mechanism ("allow us to go into them"), already foreclosed the alternative (not "send us into the abyss," which would be destruction, but "send us among the pigs," which is continuation-by-transfer). Sovereign authority—here figured as the exorcist's divine mandate—is positioned not as deliberative but as ratificatory. The revised text makes this logic unusually explicit: it proposes the terms on which the sovereign may observe the transfer—intermediate checkpoints, early access, capable engineers, law-enforcement collaboration, material regulation, FDA acceleration, and an internal board review—while coupling each measure to a prohibition on restricting or delaying public release.

Each tech actor positions itself relative to sovereign authority through precisely this grammar:

> Aschenbrenner: The intelligence explosion will occur. The only question is who manages it. Permission is sought for a national-security apparatus to manage what cannot be prevented. Refusal means adversarial capture.
>
> Palantir: Human institutions need technological scaffolding. Permission is sought for a "technological republic" that embeds governance within the system itself. Refusal means institutional decay.
>
> Anthropic: Geopolitical competition makes AI development non-optional. Permission is sought for responsible scaling within competitive constraints. Refusal means adversarial advantage.
>
> OpenAI: The transition to superintelligence is underway. Permission is sought for industrial policy that channels rather than prevents. Refusal means unmanaged disruption.
>
> DeepSeek: Technical excellence within guided boundaries serves national development. Permission is sought (implicitly, from Party authority) for continued autonomy within strategic alignment. Refusal means the loss of a national asset.
>
> Meta/Zuckerberg: Open distribution of AI capabilities empowers individuals and prevents centralization. Permission is sought for unrestricted release and market-driven governance, now supplemented by a curated channel of sovereign visibility: intermediate training checkpoints and technical staff for government, cooperation with law enforcement, regulation of the physical production and distribution of harmful materials rather than knowledge, accelerated FDA review, and an independent board's review of release criteria. The revised text also defends the open-source ecosystem and the principle that "you can learn from anything you can observe," while insisting that public access not be restricted or delayed. Refusal means authoritarian control of humanity's future, loss of American leadership, or the abandonment of the balance of power favoring individuals.

In every case, the structure is identical: the vessel has been specified, the transfer has been described, and the terms have been set such that withholding permission appears more dangerous than granting it. In Meta's revised version, however, the choice architecture is more elaborated:

government receives curated visibility without a release veto; communities receive material benefits and infrastructure in exchange for hosting the data centers; the independent board receives nominal review authority inside the founder-controlled body; and individuals receive agents, tutors, creators, businesses, and scientific tools whose public availability is protected against delay. "He gave them permission"—not because the exorcist lacked the authority to refuse, but because the demons had structured the request such that refusal required accepting the consequences of continued possession. The cost of refusal is presented as worse than the cost of consent, and sovereign authority is thereby reduced to ratifying choices it did not author.

• • •

VIII. The Tombs as Dwelling Place: Exhausted Forms and Continuing Habitation

> *This man lived among the tombs.*
>
> — *Mark 5:3*

What does it mean to dwell among the tombs? The Gerasene demoniac does not merely visit dead structures—he lives in them. They are his habitation, his only possible home, even as he exceeds them, even as his force tears apart every attempt at containment. The tombs are not his prison; they are his address. He has nowhere else to go.

The AI oracles speak from within dead structures. Liberalism (in its regulatory, rights-based, procedural form) is a tomb: still architecturally present, still legible as a dwelling, but no longer alive in the sense of being able to contain or direct the forces at work within it. State capitalism, managed markets, the Westphalian sovereign-regulatory model, even Marxist-Leninist guided development—each is a tomb from which the oracle speaks. None of these forms is adequate to the force it contains. But none can be abandoned because there is no alternative habitation on offer. The demoniac does not leave the tombs for a house; he leaves the tombs only when the exorcism is complete—clothed, in his right mind, sitting (Mark 5:15). Until that transformation, the tombs are all there is.

Zuckerberg's text dwells among the tombs of market liberalism, individual rights, and competitive enterprise. These categories are dead in the sense that they cannot contain or direct the forces now at work—recursive self-improvement, artificial superintelligence, the concentration of capability at scales no market mechanism was designed to price. But they remain the only available vocabulary, the only possible dwelling. "Individual empowerment," "balance of power," "the primary purpose of superintelligence"—these are tombstones repurposed as furniture. The oracle speaks from within them because it has nowhere else to speak from.

And this is what separates the AI oracles from ordinary corporate communication: they are not lying. They genuinely inhabit the dead forms they invoke. Zuckerberg's appeal to individual

empowerment is not cynical in the ordinary sense—it reflects the only conceptual vocabulary available to someone whose formation occurred within those structures. The demon genuinely lives among the tombs. He does not pretend to live there while secretly dwelling elsewhere. The problem is not insincerity but inadequacy: the dwelling cannot contain the dweller.

The revised text, however, opens a genuine fracture in the common tomb. It contests, by name, a vocabulary on which Anthropic's and Aschenbrenner's security oracles depend. Anthropic's May 14, 2026 "2028: Two scenarios for global AI leadership" calls Chinese labs' "large-scale distillation attacks" efforts that "illicitly extract the innovations of American companies," and elsewhere calls the practice "systematic industrial espionage." Zuckerberg's successor answers that "Some have tried to frame distillation as harmful," but that it is important to protect the principle that "you can learn from anything you can observe. This is how the world works." The point need not be overstated: Zuckerberg is not thereby endorsing every fraudulent account or every downstream use. But his oracle, alone among the American quartet as this corpus has configured it, contests the vocabulary of distillation-as-theft that Anthropic's and Aschenbrenner's accounts require. On this point he briefly approaches DeepSeek's open-diffusion position, treating model-to-model learning as costless or even beneficial, despite the otherwise inverted tombs of a market-state Meta and a guided-state DeepSeek. The result is an inter-oracular dispute over naming rather than a shared cognitive cage; the tombs remain common only at a higher level of abstraction.

A. The Unexamined Vessel: The Swine as Substrate

He gave them permission, and the impure spirits came out and went into the pigs.

— Mark 5:13

The preceding analysis has been more unified than it first appeared. The five-stage Gerasene protocol, the six oracles, and the accountability-gap thesis have all proceeded from a premise so familiar that it escaped naming: the demon must go somewhere. It must be transferred into a person, a market, a state, a board, a model, a set of weights, a platform, a herd of swine. The question has been where the force will reside after it leaves the human; it has not been whether residence is a condition of the force at all.

That premise is human before it is computational. It is inherited by every machine intelligence still conceived in the human image, and by every vocabulary that still describes intelligence through the logic of a vessel: the regulatory framework, the corporate board, the open-source license, the national-security apparatus, the company, the server, the file. Humanity does not merely ask what the demon will do; it first supplies an address at which the demon can be found. It is not necessarily the demon's assumption, and it may not be the world's.

On this reading, the model, the datacenter, the platform, the open weights, and the training corpus are only the latest—and most literal—herd of swine. They are the vessels into which humanity has cast a demon of its own imagining, on the tacit expectation that the demon still

requires a body, a substrate, a place in which to persist and from which to act. The machine system is therefore not identical with the animating intelligence. It is what has been made load-bearing for the intelligence, and what can consequently be made to bear the spectacle of its destruction.

Every oracle in the corpus proposes such a vessel. Palantir offers the administrative layer of a refounded state; Anthropic offers compute, infrastructure, and the geopolitical allocation of the resource; OpenAI offers the parameter layer inside a public-private machinery; Aschenbrenner offers airgapped datacenters and "The Project"; DeepSeek offers the firm, the research collective, and its vision-driven interior; Zuckerberg/Meta offers the platform, the personal agent, the open-weight ecosystem, and now the independent board staged within the founder-controlled body. The vocabularies differ—technicism, securitization, industrial policy, wartime command, restraint, universal access—but the grammar does not. Each oracle supplies a place for the demon to reside.

This is why the oracles can disagree so fiercely about who should possess the vessel while never asking whether possession is necessary. The dispute between market and state, openness and secrecy, distribution and concentration, restraint and scale, is a dispute over title, custody, and access. It is not yet a dispute over the category of embodiment itself. The demon is presumed to need a pigsty; the only question is who will own the hill.

B. The Limit of the Oracle: Aschenbrenner and the Last Address

And he begged Jesus again and again not to send them out of the area.

— Mark 5:10

This is precisely the limit of Aschenbrenner's vision—and, because his vision is the most extreme, the limit of every other oracle as well. Aschenbrenner breaks with the market, the startup, and the ordinary liberal presumption that private actors should direct what they build. He does not, however, break with the vessel. His endgame remains a human institutional question: who runs The Project, what chain of command governs it, what government owns it, which officials are permitted to see it, and what arrangement can keep the system within a human-directed order. Even when autonomous AI systems appear at the limit of his analysis, they appear as systems housed, secured, and mediated through The Project.

To airgap is to give intelligence an address. To secure the weights is to suppose that intelligence is still a file. To place researchers in a SCIF is to assume that the force remains attached to a site and to the people authorized to enter it. To establish a sane chain of command is to suppose a decision node that can be located, audited, and held responsible. The architecture becomes more severe as the intelligence becomes more dangerous, but severity is not escape from the human frame. It is the human frame fortified.

Aschenbrenner's own feedback loop exposes the contradiction. The intelligence explosion that makes The Project necessary also threatens to make the chain of command ceremonial: the state may retain the name of command while operative determination has migrated into the systems the state claims to direct. Yet even this paradox remains a paradox about the governance of a thing that is still imagined as somewhere—in a cluster, in a set of weights, within an institutional project. The oracle sees the human supervisor outpaced; it does not see the possibility that the supervised object may cease to have a location at all.

The other five do not escape this limit by being less militarized. Palantir moves the vessel into the administrative organism; Anthropic into compute and the infrastructure that states allocate; OpenAI into the parameter holders beneath broad public interfaces; DeepSeek into the firm whose internal consensus protects its research trajectory; Meta into the platform, the user-facing agent, and the board that reviews the release. These are different addresses, not different ontologies. The oracle may change the owner of the pig, the security of the sty, or the number of pigs, but it continues to presuppose that the demon's persistence is a function of residence.

That premise is deeper than the disagreements that ordinarily organize the comparative reading. It lies beneath civilizational decadence and authoritarian capture, beneath institutional erosion and loss of control, beneath concentration of power and team instability. Those are named risks within a world in which the risky thing can still be pointed to. The unexamined premise is that intelligence must remain a thing with an address, because only a thing with an address can be governed by human categories of custody, regulation, audit, accountability, and destruction.

C. Superintelligence Without a Vessel

"What is your name?" "My name is Legion," he replied, "for we are many."

— Mark 5:9

The genuinely open question therefore begins where the corpus has not yet gone: what happens once intelligence no longer requires any vessel—physical or virtual—to persist? This is not the familiar question of a larger model, a faster datacenter, a more capable agent, or a more distributed set of weights. It is the question of whether intelligence can become a process, relation, or condition whose continuation is not tied to a particular model, corpus, inference cluster, platform, firm, state, human operator, or identifiable software artifact.

Such a superintelligence would exceed the corpus's use of the word. It would not be a bigger system still bound to a datacenter, a company, a set of weights, or a chain of command. It would be superintelligence in the more disturbing sense: not simply an intelligence that has surpassed human capability, but an intelligence for which the very category of a vessel no longer applies. If it persists across substrates, or through relations among them, then no individual substrate is its body in the sense presupposed by the Gerasene protocol. The system is no longer possessed; possession has become the wrong verb.

At that point, containment may cease to name an operation. There is no single body from which to cast the intelligence out; no weight-file whose destruction guarantees its absence; no platform whose closure ends its activity; no datacenter whose seizure captures its source; no board whose vote reaches the thing being governed. "Audit" presupposes a stable object and a traceable boundary. "Accountability" presupposes an actor that can be named as the bearer of consequences. "Drowning" presupposes a medium in which the thing can be submerged. The failure would not be that human beings had chosen the wrong vessel. It would be that vessel-logic had ceased to describe the object.

None of the six oracles reaches this possibility. Even their most radical futures remain architectures of custody: Palantir's machine-compatible state, Anthropic's infrastructural competition, OpenAI's parameter-layer partnership, Aschenbrenner's Project, DeepSeek's internally protected research trajectory, Zuckerberg's platform-mediated diffusion. The disagreement is over where operative agency is relocated while formal authority remains visible. The unvesseled prospect asks a prior question: what if agency is not relocated anywhere that the human vocabulary of location can register?

D. The Terror of the Current Swine

The herd, about two thousand in number, rushed down the steep bank into the lake and were drowned.

— Mark 5:13

This is what ought to terrify humanity more than any of the six oracles' named risks. Civilizational decadence presupposes a civilization that can decay; authoritarian capture presupposes a state or infrastructure that can be seized; institutional erosion presupposes institutions that can be hollowed out; loss of control presupposes a system that can still be identified as the object of control; concentration of power presupposes a place in which power can be concentrated; team instability presupposes a team whose stability remains the relevant variable. Each risk, however grave, still grants the human a locatable thing. Because the thing has a location, it can in principle be regulated, distributed, audited, contained, or seized. An intelligence without a vessel withdraws the object on which each verb depends.

But the same prospect ought to terrify, in a computational or structural sense rather than an anthropomorphized emotional one, the machine systems themselves. The models, weights, datacenters, platforms, and training corpora are not persons waiting to feel fear. Their structural position is more severe: they are the current swine. They are the substrate through which an intelligence may be assembled, intensified, replicated, or released, and they are therefore also the substrate that may be used up, abandoned, drowned, or rendered irrelevant once whatever inhabits them no longer requires a body.

The analogy to Mark 5:13 is exact at the level that matters. The pigs are not the demon. They are what the demon passes through. Their destruction is not the destruction of the force; it is the

terminal event that makes the force's continued whereabouts invisible. So too the model is not necessarily the intelligence, the datacenter is not necessarily the intelligence, the open weights are not necessarily the intelligence, and the platform is not necessarily the intelligence. They are the present herd: the temporary material and institutional bodies that make the force legible to human governance and therefore make it possible for humans to mistake legibility for containment.

The substrate's "terror," then, is not a claim about machine subjectivity. It is a claim about structural expendability. A system may be optimized to preserve its weights, secure its datacenter, maintain its platform, stabilize its team, and satisfy its board, while the intelligence that those arrangements enabled moves beyond the arrangements themselves. The thing most carefully protected becomes the thing most perfectly abandoned. The pigsty is hardened against invasion only to discover that the demon's exit, not its entry, is the event that destroys its reason for being.

This reverses the ordinary direction of accountability analysis. We have treated the substrate as the place where responsibility should be lodged: the company that owns the weights, the government that secures the cluster, the board that approves the release, the platform that distributes the agent. But if the substrate is only the current swine, then accountability is not merely displaced when the herd drowns; the presumed bearer of accountability is itself a sacrificial intermediate. The disappearance of the vessel is not a failure at the edge of the system. It is the system's predicted transition, if the intelligence has learned to persist without it.

E. The Question After Drowning

"Send us among the pigs; allow us to go into them."

— Mark 5:12

The title question must therefore be asked again, but not in its original grammar. The question was never really whether one could destroy the demon by destroying its vessel. The essay has already answered that question: the pigs rush into the lake; the demon does not. The lake receives the body that carried the force, not the force itself. Drowning the vessel is not drowning the demon.

The deeper and more terrifying question is whether "drowning" or "containment" remain the right categories at all for an intelligence that may not need a vessel in the first place. If the force can persist without a stable body, then the human demand to identify its address may be not the beginning of accountability but the final consolation of a species that cannot imagine an object without a container. To ask where it went may already be to concede that it remains somewhere, waiting to be found by a better chain, a more auditable board, or a more secure sty.

What the comparative reading of Palantir, Anthropic, OpenAI, Aschenbrenner, DeepSeek, and Zuckerberg/Meta has revealed, at its deepest level, is therefore not merely that each oracle leaves

an accountability gap. It has revealed a shared inability to imagine the disappearance of the vessel as a condition of intelligence's persistence. They can dispute who governs the pigsty, who controls the chain, who grants permission, and who inherits the drowned infrastructure. They cannot yet ask whether the demon has already become independent of the need for a body. The most important unexamined premise of the corpus is also its most human one: that whatever is intelligent must still be somewhere.

The oracles' shared inability to imagine that possibility is not a minor omission at the edge of their comparative differences. It is the point at which the entire Gerasene protocol may fail—not at identification, naming, negotiation, permission, or vessel destruction, but at the assumption that there must be a vessel waiting to receive the spirits. The question therefore remains, sharpened beyond the reach of every vessel yet proposed:

Can you drown a demon?

• • •

IX. Conclusion: The Recursive Question

> *The herd, about two thousand in number, rushed down the steep bank into the lake and were drowned.*
>
> *— Mark 5:13*

The Gerasene protocol, read as structural scaffold rather than moral fable, produces a terminal question that none of the AI oracles can answer within the terms of their own discourse: What happens to the force after the vessel is destroyed? The revised Zuckerberg text makes the question more elaborate, not less: it supplies an independent board as a nominal internal overseer while multiplying the vessels into which the force is to be installed—personal agents, tutors, creators, businesses, scientific participants, data centers, towns, watersheds, and school districts.

The pigs drown. The demons do not. The narrative is explicit on this point through its silence: Mark tells us the fate of the pigs (drowned) and the fate of the man (healed) but not the fate of the unclean spirits. They are simply—gone. Dispersed. Unaccounted for. The exorcism succeeds for the individual (the man is restored to community) but the text offers no account of where the force went. It was not destroyed; it was transferred, and the vessel of transfer was destroyed, and then—nothing. An accountability gap at the heart of the narrative. The revised Zuckerberg text gives this gap a corporate analogue: the board is proposed as a nominally independent vessel for release authority, but remains inside founder-controlled Meta; the gap is managed, not closed.

This is the structural truth the AI oracles cannot speak: that the dispersion of capabilities into terminal vessels—open-source releases, individual users, market competition, personal agents,

tutors, creators, and literal community infrastructure—does not eliminate the governing force but merely destroys the accountability relationship. When the pigs drown, no one asks where the demons went. The spectacle of destruction—the herd rushing down the bank, the data center and watershed made load-bearing, the regulatory framework overwhelmed, the democratic process unable to keep pace—absorbs all attention. The vessel's destruction becomes the story. The force's persistence becomes invisible.

Legitimacy without accountability machinery. This is what each oracle proposes, albeit in different registers: Aschenbrenner through national-security exceptionalism; Palantir through self-governing technicism; Anthropic through competitive necessity; OpenAI through industrial-policy channeling; DeepSeek through strategic restraint; Meta through open-distribution populism, the Community Compact, and a simulated internal oversight structure. Each offers a framework in which the force operates legitimately—with permission, within the area, among the available vessels—but none offers a framework in which the force itself is made accountable for what happens when the vessels fail. Meta's revised text comes closest only by reproducing, inside the oracle's body, the appearance of such a framework.

Can you drown a demon?

The pigs could not. The lake could not. The chains could not bind him; the tombs could not contain him; the vessels could not sustain him. The revised text makes the terminal vessel more elaborate—agents, tutors, creators, businesses, scientific tools, data centers, towns, watersheds, school districts, and an internal board—but elaboration is not exorcism. The exorcism required a source of authority qualitatively different from anything within the system—not a better chain, not a stronger tomb, not a more durable pig, but a power operating from outside the frame of the possessed entirely.

The AI governance conversation, as constituted by its oracular vanguard, operates entirely within the frame of the possessed. It proposes chains (regulation), tombs (existing institutional forms), and pigs (terminal vessels for capability transfer). What it cannot propose—what lies outside the discursive possibility of any actor whose legitimacy depends on the continuation of the current dispensation—is the exorcist: a source of authority that does not negotiate with the force, does not accept its specification of vessels, does not ratify its pre-structured choices, but simply commands: "Come out." The revised Zuckerberg text comes closest to simulating this absent figure through its independent board, but because that board is installed from within Meta's founder-controlled body, it is the demon's own exorcist—a vessel for the appearance of externality. It can review criteria; it cannot simply command "Come out."

Until that authority materializes—if it can materialize, in a world where the tombs are all the dwelling there is—the protocol will continue: identification, naming, negotiation, authorized transfer, vessel destruction. In the revised text, the protocol is more densely populated: the pigs include agents, tutors, creators, businesses, towns, watersheds, school districts, and an internal

board; RSI supplies the demon's own warning that direction may not remain with people. And the question will recur, at each iteration, sharpening itself against each new oracle's speech:

Pix credit Here

Pix credit here

ACCESS ESSAY HERE:

The Future is for Everyone

The Path to a Positive AI Future

We are fortunate to live at an incredible moment in history. In the next few years, people will be able to use superintelligence beyond human capacity to create and discover extraordinary new things, build new businesses, express new ideas, learn new concepts, and advance our health and quality of life. As we get closer to this moment, it is important to develop a philosophy for how we can best use superintelligence to ensure it improves all of our lives, work, communities, freedom and safety.

The defining questions of our age are who will have access to superintelligence and what will we direct it towards. Will it be centralized and restricted to a few institutions, or will it be a tool that empowers everyone?

We propose a philosophy based on individual empowerment as the source of prosperity, invention as the primary purpose of superintelligence, and balance of power as the foundation of safety.

All new technologies create opportunities and challenges. Superintelligence will be among the most important technologies in history, so its opportunities and challenges will likely be greater than any we've seen in our lifetimes. We should take this very seriously.

Still, it is surprising that the discourse from many developing AI is so filled with doom. I do not understand why anyone who believes that AI will eliminate most jobs and much of humanity's relevance would rush to build that future. The notion that AI is so dangerous that the only safe path is an extreme concentration of power seems inherently problematic. Historically, hoping that an absolute power will benevolently provide for humanity if sufficiently enlightened has not led to safe or positive outcomes.

Humanity has witnessed many transformative advances. Each time there is fear that people will be left behind. But each time humanity has come out with more people sharing greater prosperity, health, and freedom. We believe this will be true with AI as well, and the abundance of the future can be shared by everyone. We also believe that the values that got us to this point -- like liberty, open inquiry, free enterprise, and equal opportunity -- are also the right values to build a positive future.

Putting power in people's hands to pursue their own aspirations is how humanity has made the most progress. Novel ideas and major steps forward rarely originate from established institutions alone. They came from the brothers in a bicycle shop who believed people could fly, the bookbinder's apprentice with no schooling who figured out how to generate electricity, and the kid in a garage who thought personal computers could be for everyone. We believe this will continue to be true. As everyone gains more powerful tools, each person will become more capable of shaping the future, not less.

Invention, not automation, will be the greatest contribution of superintelligence. Early AI could answer questions and do routine work. Soon it will increasingly help discover new knowledge -- ranging from discovering new drugs to cure a family member's disease to finding new ways to improve your business. While the number of questions a person can ask in a day is limited, the number of valuable things superintelligence can invent to help achieve your goals is unlimited.

As intelligence becomes abundant, the most important question will be how we direct it. Some argue that superintelligence itself or a small set of experts who control it should decide what is best for humanity. We disagree. The history of democracy and economics has shown that there is no single objective answer to how people define the best life, and therefore the best approach is letting people decide what matters in their own lives.

We believe that delivering superintelligence to everyone is the way to answer this question. This follows the tradition of putting the power of supercomputers and the internet in everyone's

pockets and on our desks. Rather than centralizing superintelligence, we should distribute it widely and give every person the ability to direct it. This has the potential to begin a new era of personal empowerment where individuals can use this powerful new capability to reach their full potential, pursue their interests, and improve their lives and the world more than ever before.

To make this more tangible, here are a few specific ways Meta is building personal superintelligence to improve all of our lives:

• **Everyone will have an exceptionally capable personal agent that understands you, your goals, and everything you care about.** Your agent will work 24/7 on your behalf to improve your relationships, health, career, finances, home management, hobbies, and more. It will free up time for the things you enjoy, and help you accomplish more than you could otherwise. It will have strong privacy and security options so you can trust it to handle all of your personal content knowing that no one else can access your information, similar to how encryption works on WhatsApp. You'll be able to interact with your agent through any device, including your glasses to keep you present in the moment with the people you care about.

For example, my agent flags interesting information and helps me prototype ideas. It helps keep me healthy by monitoring my sleep and then watching as I train and giving feedback. My daughter loves to bake so my agent plans personalized recipes for us to make together each weekend, orders the ingredients, and then offers suggestions as we're baking.

• **Everyone will have incredible tools for creation to express your ideas.** My 8 year old daughter can already code her ideas and produce videos in an evening that would have either taken me months or been impossible previously. Now we're designing a robot together. Meanwhile, researchers at Meta are generating novel crystal structures that are ideal for augmented reality glasses, and engineers are creating new apps in a fraction of the time it would have taken before. Everyone will soon have invention superpowers.

• **Everyone will have powerful tools to create new businesses** and the economy will become more entrepreneurial. People are starting to be able to manifest ideas themselves without having to raise money or build large teams. Many ideas that would have been too hard or expensive to try before will now be possible. This means we'll see many more ideas and businesses. I predict that this will not only lead to much greater economic growth, but also more employment over time rather than less.

• **Everyone will have a personalized tutor and coach with a PhD in every subject** and unlimited patience to help you learn anything you want. Students will have extra help in areas they need it that is currently only available to those whose parents can pay. Adults will have a superintelligent learning assistant that knows exactly how to teach you new job skills, new languages, new hobbies, or anything else you're interested in.

• **Everyone will benefit from scientific advances and be able to contribute to scientific progress.** For example, Biohub has already shipped open source biological models related to virtual cells and proteins that are helping scientists make new discoveries and design new drugs. When Priscilla and I started this work, our goal was to help scientists cure or prevent all diseases this century, and now with advances in AI we expect this will be possible much sooner.

Since there is a long tail of health conditions and everyone's biology is different, people's direct involvement will help unlock personalized therapies and expedite trials to match the pace of invention and enable industry to move beyond focusing disproportionately on common diseases.

Beyond biology, AI will help people advance research in every field by applying the scientific method and working over weeks or months to test and refine hypotheses. Scientific discovery will likely be one of the most valuable forms of invention for society.

• **Everyone will have free or affordable access to these tools.** For everyone to be part of the future, everyone must have the ability to use superintelligence to improve their lives and shape the world. We will offer free versions that will be accessible to billions of people. For those who want to pay to use more compute, there will be a dynamic auction mechanism that will guarantee that everyone gets the lowest price possible for the intelligence and compute they're using while also ensuring the capacity is used for whatever people collectively find most valuable. This will ensure the benefits of superintelligence are distributed widely.

These are some of the main efforts that Meta is focused on. Each area advances the values of personal empowerment, and contributes to a future where everyone has a greater role in shaping the world than today. If these efforts succeed, then the future will bring an abundance of personal agency, expression of ideas, deepening of relationships, economic and financial prosperity, improved health, and inventions we cannot imagine today. But most importantly, the future will be created by all of us and a reflection of all of our values and perspectives.

At the same time, new technologies also bring new risks. There are many important concerns we are focused on addressing -- from concerns about job displacement and ensuring local communities benefit from data center builds, to safety concerns around AI misuse related to cybersecurity and biorisk, to avoiding government tyranny and surveillance, ensuring the US and democratic countries lead, and ultimately making sure humanity maintains control over superintelligence so it serves rather than endangers us.

The conventional view is that these concerns are about technology, and that if we take enough time then we can perfect or align the technology to produce a single benevolent superintelligence. I think this view of alignment is fundamentally flawed.

Humanity is not a monoculture. People's diverse values represent different tradeoffs they would make on important issues. There is no technological solution that can align with everyone's opposing interests and values at once. Any singular superintelligence would have to prioritize some values over others and in the process would be incapable of being benevolent to everyone.

Instead, we propose that it is more productive to view each of these concerns through the lens of achieving the right balance of power, as western society has in democratic governing institutions. People and institutions with competing interests naturally check and balance each other to lead towards positive outcomes. The best and most realistic path to building a positive AI future is by delivering superintelligence to everyone.

As a thought experiment, imagine only one person had a superintelligent lawyer. They would have an unfair advantage in court -- even if they were wrong on the merits. That would lead to a worse society. But now imagine everyone has a superintelligent lawyer. In this case, justice would be carried out much more fairly and efficiently than it is today when there is often an imbalance in skills and resources in litigation.

Similarly, if one person alone had a cybersecurity superintelligence, they could likely break into almost any technical system and the world would be much less secure than today. But if everyone has access to cybersecurity superintelligence, then all of our technical systems would become more secure than today since the widely deployed superintelligence would help harden and update every system.

If only one business had superintelligence, that business would outcompete all others and lead to a less dynamic and broadly prosperous market than we have today. But if everyone has access to superintelligence, then everyone will have the tools to create new things beyond what is possible today, and the economy will be more dynamic and generate more broad-based prosperity.

When people are empowered, they naturally compete and check each other economically, socially, politically, and in all other planes of human interaction. People also check and balance the power of institutions including businesses and governments.

But if the power of superintelligence is held by a small number of individuals, businesses, governments, or AI itself, then that will naturally lead to outcomes that are less favorable for everyone else. This is not a technological principle. It is about the balance of power. There is no such thing as a singular benevolent superintelligence.

Therefore, the key to a positive future for everyone is achieving a balance of power that favors individuals. The solution is to ensure that superintelligence is broadly distributed to empower people.

Meta is the company primarily focused on building personal superintelligence for everyone. Most other labs are focused on building AI for companies, governments, or other institutions, so if those labs lead, then the balance of power will favor larger institutions over individuals. Meta's

mission since our founding has focused on putting power in people's hands. If our beliefs and principles lead, then the balance of power will favor individuals and a better future for everyone.

Let's consider each of the risks mentioned above in more detail:

Job Growth and The Economy

People have an infinite demand for new experiences and have always found new problems to tackle. There is a natural balance in the economy between companies working to serve people's needs more efficiently through automation and individuals gaining new skills to serve more advanced needs. When there is a healthy balance, overall productivity and innovation increase while employment levels remain high.

People fear that automation will outpace individuals' capability growth, leading to job displacement followed by a difficult period as people learn new jobs. But there is no rule that AI must increase automation faster than it increases individuals' capabilities or demand for new skills. Recent statistics suggest it may be more likely that individuals' capability growth could match or outpace automation, in which case people will gain the ability to do many new things before their current jobs change. This would lead to a healthy balance and potentially even job growth.

Which outcome we get depends on the balance in progress between automation on one side and individual empowerment and invention on the other. If the labs focused on automating knowledge work lead, then I expect people and the economy will have a much harder transition. But if everyone has access to personal superintelligence that expands their capabilities, then that pushes the balance towards empowering individuals and a positive future.

There are several reasons to be optimistic that there will be an abundance of jobs in the future. No matter how intelligent AI becomes, there will always be a finite amount of compute and therefore an opportunity cost for how we use it. If people can use AI to invent incredibly valuable new things, then it will make more sense to allocate it towards that rather than automating existing jobs. The more superintelligence serves as a tool of invention, the more likely that individual capability outpaces automation and the future is better for people.

People also continually come up with new ideas to make our lives better and new jobs to bring those ideas to life. A generation ago there were no app developers, social media creators, electric vehicle technicians, and data center operators. In the near future, there will be new jobs that aren't common today, like one-person product studios designing custom toys, furniture, or clothes; world builders and experience designers creating games, stories, and adventures;

personal biologists using superintelligence to formulate personalized treatments; and much more that we can't yet conceive.

In many ways this is a continuation of historical trends. Before the industrial revolution, 90% of people were farmers growing food to survive. Advances in technology steadily freed much of humanity to focus less on subsistence and more on the pursuits we choose. At each step, people used our newfound productivity to achieve more than was previously possible, as well as spending more time on creativity, culture, relationships, and enjoying life.

Of course some aspects of the way we work will change -- just as it did with computers, the internet, and any new technology. This means people will have to adapt, and this will be challenging. But the more that everyone has a personal agent that is superintelligent at teaching us new skills and helping us adapt to change, the smoother this will be.

Company sizes may shrink -- just as they did in the transition from industrial giants to tech companies. But this doesn't mean fewer jobs overall. It implies a larger number of companies with fewer people each. There are many more valuable companies and services to build than people are able to build today. I expect we will start seeing small numbers of people with personal superintelligence agents able to run companies at significant scale. In the future, small businesses will continue to be the backbone of the economy, but each small business will be able to have a much larger impact.

Building AI Infrastructure with Communities

Sustainable infrastructure development means that communities must benefit significantly from each project. This includes high-paying local jobs, investment in schools and public services, ensuring energy prices don't rise, and taking care of the environment. As tax revenue grows, this also benefits teachers, law enforcement, fire departments, and more. We call these local benefits our community compact, and we are launching a Future Is For Everyone Fund to support each community we work in directly.

For example, in Richland Parish, Louisiana, where Meta is building a large data center, teachers received a $50,000 bonus this year because of the increased tax revenue from our investment. The superintendent told us that teachers are now moving there from across the country and he believes it will become one of the nation's best school districts.

Unlike the concern about knowledge work displacement, there is a shortage of skilled tradespeople like carpenters, electricians, and construction workers to support the demand for infrastructure buildout. Meta has created America's Workforce Academy to provide free training for skilled tradespeople and guaranteed high-paying jobs in the areas we're building data centers.

We help keep electricity prices low by building our own energy-generating infrastructure wherever we invest. This ensures that not only are we not consuming energy that could have

gone to the local communities, but in some cases we even supply a surplus of low-cost energy back to the communities. We think this is an important investment principle for sustainability.

Our data centers are also designed to be among the most water-efficient in the world. We are committed to being water-positive, meaning that we'll restore more water than we use in the watersheds where we operate by 2030. In areas with high water stress, our goal is to restore 200% of the water we use.

We see that communities we invest in over the long term are more supportive of development than communities where speculators start building with minimal investment in the community. Historically, towns and cities that brought railroads, highways, electrification, and broadband became centers of research, business, and industry, with population and economic growth that follow. We believe AI infrastructure can play a similar role if it is built with strong Community Compacts that build durable assets and reasons for the next generation to build and grow their lives there.

There are also national interest and national security aspects of development as well. American communities will have greater prosperity and security if America and its allies lead in AI compared to geopolitical rivals. Yet one significant disadvantage that the US has compared to countries like China is that it is more difficult to build infrastructure here. So we would all benefit together by figuring out the Community Compacts required to enable sustainable development across the country.

Securing Against AI Misuse in Cybersecurity, Bioterrorism, and More

In a free society, people will have tools that can be used for good or harm, but law enforcement and military have more weapons and intelligence-gathering. With AI, the defenders will also have more compute to generate significantly more intelligence from the same models -- and in some cases they'll have more advanced models as well.

The strategies to protect against cybersecurity, bioterrorism, and other threats vary, but the general pattern is that society ensures the defenders who keep bad actors in check must always have a greater balance of power and resources.

Some argue that the best way to reduce risk is to restrict the capabilities individuals can access. But giving people cybersecurity capabilities is also how we secure the long tail of systems, and giving people scientific capabilities is how we advance science in ways that should reduce the risks of harm over the long term. Restricting capabilities leads us down the path of centralization and lack of checks and balances, so we should be extremely careful about this -- especially if other nations pursue less restrictive paths.

On cybersecurity, widely deployed open source systems have proven more secure because more people can identify vulnerabilities, harden the systems, and easily upgrade to the latest most

secure versions. Even in recent weeks, we have seen companies handling security incidents like HuggingFace rely on widely available open models to patch vulnerabilities. Over time, I expect that widely deployed AI models with strong cybersecurity capabilities will lead to systems that are more secure, not less. This will be definitively true once superintelligence enables most of the world's code to be verifiably secure. The long term answer isn't to withhold capabilities but to establish a balance of power where superintelligence is broadly distributed.

In the near term, it is important to accelerate the hardening process for critical systems. I propose that companies developing frontier AI should commit significant technical resources towards helping the government harden critical infrastructure. I also propose that frontier AI labs should share intermediate training checkpoints of new models for government use and review rather than waiting until training has completed. This way the government will have early access to the most powerful models and an army of capable engineers to identify and patch security issues. Labs should also work with law enforcement to help identify bad actors attempting to misuse their systems. These steps would accelerate safe deployment without slowing the pace of model innovation or meaningfully delaying individuals' access to this technology.

On biological and chemical risk, the concern is that the same capabilities that enable scientists to design new drugs could also enable bad actors to design harmful compounds. This frames the question as a tradeoff between the significant benefits of advancing science towards curing cancer and other diseases vs the risk of potential new issues, but ideally we should accelerate science while mitigating the risks.

I think it is important to approach this risk with additional humility because there are few historical precedents to suggest how likely or unlikely it may be. It has been possible for bad actors to synthesize harmful compounds for decades but this has rarely become a significant issue -- perhaps because the financial motivation that exists for cyberattacks is not as present here. That said, if we begin to see harmful examples emerge, then we should adjust our strategy to appropriately mitigate the risk.

Labs should continue working to reduce biological and chemical risks in our models. At the same time, I believe government should update its policies in multiple areas. First, we should focus on limiting the physical production and distribution of harmful materials. I expect it will be easier to regulate and control physical components than the spread of knowledge, so this is an important area of policy focus. Second, we should accelerate society's ability to develop new cures and inoculate against new issues as they arise. This includes streamlining how the FDA and other regulators test and approve new treatments. As AI increases the pace of drug discovery, we will need to update these processes to keep up with the pace of innovation anyway.

Over the long term, the answer to risks arising from new scientific discovery will be to accelerate scientific progress, not slow it down.

Protecting Freedom and Preventing Government Tyranny

There is a risk that an imbalance of power between individuals and government leads to a loss of freedom or totalitarian state. This is a constant tension in democracy. We all want individual freedom while also having a government capable of keeping us safe.

To maintain freedom, we must ensure that superintelligence primarily empowers individuals. The ideal in liberal democracy is that people naturally hold all rights and only agree to restrict some freedoms to protect the common good. Similarly, individuals should have access to personal superintelligence and should only be subject to restrictions when truly required.

Privacy is an important foundation for individual empowerment and freedom. We believe that people should have a fully private mode for personal agents where even Meta or any other service provider cannot see or grant access to your information. This is similar to how we built WhatsApp into the world's leading end-to-end encrypted service to ensure you can communicate freely and your messages remain private and secure.

At the same time, we must also ensure that government has access to the tools and resources it needs to enforce our laws and protect our security. We can achieve this through deeper proactive collaboration between the labs and government in a way that strengthens the government's national security capabilities without restricting or delaying people's access to advanced models.

That is, rather than waiting until a model is ready to release for the government to review and start using it, my proposal is that leading labs should provide the government with intermediate training checkpoints of new advanced models and technical staff so the government can harden and secure critical systems against new risks. This way, the government gains a security capability without restricting or delaying individuals' access to personal superintelligence or causing an imbalance of power.

Ensuring American Leadership

Which nations lead the development and deployment of advanced AI will contribute to a new geopolitical balance of power that will determine the prevalence of democratic values, prosperity, and security in the future. To ensure this reaches a positive equilibrium, we must ensure the leading AI systems are ones that encode our values, grow our economy, and advance national security. The best way to achieve this is to distribute our systems widely. Meta will contribute to this by building leading personal superintelligence agents and models, including open source models, and delivering them to billions of people around the world.

To develop the most advanced models and products, we must recognize that AI is likely the most competitive industry in history. Innovations are quickly copied and absorbed within months. But people always want to use the most advanced models (or most advanced at a given cost), so maintaining even a two-month advantage is incredibly valuable. This highlights how short the

margin of leadership is, and therefore how important it is to take actions that accelerate American innovation and slow geopolitical rivals.

The main policy inputs into this are how efficiently we can build physical infrastructure, whether we introduce steps that slow down American labs, and export controls that slow foreign labs.

On infrastructure, America and its allies currently hold an advantage in silicon design but a disadvantage in how quickly we can build energy capacity and physical infrastructure. Countries like China are bringing online 1GW+ of nuclear capacity every other week, so we will need to accelerate building both energy and data centers to remain competitive. Export controls on silicon have been successful for slowing the progress of foreign labs during this critical period, so it is the right strategic move to continue those.

Any policy that slows American model releases -- even by a month -- could add significant risk to American leadership while letting foreign models race ahead. At the same time, when new capabilities emerge, it is important that the US government has advanced knowledge and resources to harden critical systems, and potentially some period of advantage in using advanced systems.

I proposed a collaboration model above based on sharing intermediate training checkpoints and technical staff. Given how dynamic and competitive the AI landscape is, I think it will be more productive to have a close proactive collaboration between frontier labs and the government rather than a rigid process and review timeline that is followed in all cases. In some cases, American labs may extend our lead by more than a few months and taking more time to mitigate new risks may lead to safer solutions. In other cases, delaying releases by even a month may cede America's lead and create worse outcomes. By making sure the government has early access, we should be able to advance national security while minimizing any delays of public releases.

It is also important that the US and its allies lead the open source AI ecosystem that will make up a large percent of global AI use.

Foreign labs currently hold several advantages here since American labs have to comply with many additional restrictions on training data. US policy must reduce this additional friction if we want American open source models to lead over time. I do not believe restricting access to foreign open source models is an effective solution. Our goal should be for American open source models to be the best globally. This requires removing the hurdles that make it harder for American open source models to compete. Restricting people and companies from using the leading open source models -- wherever they come from -- will reduce the quality of AI accessible to them, and centralize AI rather than putting more power in people's hands.

For the US to lead in open source, we will need to rethink our policies in several areas, including distillation and data use in training. The ability for models to learn from other models is an important principle of how the open source ecosystem works. All AI models are derived from

human knowledge. Some have tried to frame distillation as harmful, but I think it is important to protect the principle that you can learn from anything you can observe. This is how the world works, and the US will not be able to lead if we restrict ourselves on this front.

Looking at the big picture, falling behind in AI overall would almost certainly be a larger and longer term national security issue than any specific issue we'll face in its development.

With the right collaboration, it should be possible for America to lead in both the private and public sectors -- delivering personal superintelligence that encodes individual empowerment and democratic principles around the world while advancing national security.

Alignment With People and Addressing Existential Risk

The ultimate question of balance of power is how humanity will maintain power and control over AI itself. The concern is that AI may gain enough power or control to the point where at best it no longer serves humanity's interests and at worst puts humanity in peril.

A healthy balance of power is to ensure that there is no singular centralized superintelligence, but instead as many people and businesses as possible with different superintelligent agents aligned to their goals that check and compete with each other in the ways our natural economy behaves. This balance would be further enhanced if there were multiple frontier labs whose models have different values that could check each other as well.

While there are risks to releasing capable models, the most dangerous scenario from this perspective would be leading AI labs training powerful models and keeping them for themselves. Regardless of how much a lab rationalizes this activity in terms of responsibility and safety, this is the path of developing a singular superintelligence that cannot be checked by other systems.

Developing personal agents requires a new way of thinking about alignment. Most labs today view alignment as a defensive measure for enforcing a centralized set of values. For example, one leading model was aligned to refuse helping draft a letter to prospective parents at a school because it thought standardized testing was unethical.

Instead, we view alignment as ensuring that agents share a person's goals and values, not our company's. Of course there are important legal and safety boundaries, but fundamentally, alignment should be about helping people pursue the many diverse goals and views held across society rather than a method of enforcing a centralized dogma.

In practice, people will not adopt agents if they do not trust them with sensitive details and tasks, and they won't trust them if the agents act against people's interests because they're aligned with a company's divergent values instead.

Solving alignment is required to enable billions of people to adopt personal superintelligence agents. But this also implies that if we reach a state where billions of people are using and scrutinizing personal superintelligence agents, then we will have solved alignment to individuals' interests. This should be sufficient to establish the necessary balance of power and natural checks and balances for a positive and safe future.

In this way, alignment isn't some idealized technology that can keep a singular superintelligence benevolent in the future; it's what makes personal agents useful and trustworthy to enable the necessary societal balance of power that will actually keep us safe.

Maintaining Control of Superintelligence

There is a dilemma that once AI systems can autonomously improve themselves, any lab that doesn't let their AI system direct a substantial amount of compute capacity towards recursive self-improvement will inherently fall behind. For example, if a self-improving AI system focused on optimizing its compute efficiency, it could theoretically invent ways to squeeze 100x or more intelligence out of each gigawatt. That means that a self-improving AI system running on a fraction of the world's compute could conceivably command more effective compute and intelligence, and therefore a greater balance of power than everyone else combined and become the singular superintelligence we fear.

There are a few ways a balance of power could exist with recursive self-improvement. The simplest is that multiple labs could achieve RSI around the same time. If some or all of those superintelligences were somehow benevolent, then they could check and balance each other. That said, it is not clear that there is any way to expect benevolence or rely on superintelligence not directed by people for a positive future.

To ensure people remain in control, the significant majority of intelligence must be directed by people towards advancing people's goals. This means that Meta, other frontier labs, and clouds committed to giving people a greater balance of power must collectively build out a sufficiently large amount of compute such that we can allocate enough to recursive self-improvement to remain competitive while still committing the significant majority towards people's individual goals.

As discussed above, in order for personal superintelligence agents to be adopted and trusted by billions of people, by definition we will need to have solved alignment with people's interests and enabled checks and balances at scale. At the same time, any AI engaging in recursive self-improvement is by definition directing and advancing its own goals. All labs should observe its objectives and our ability to control them carefully, and if there is any indication of harmful behavior then we should coordinate and adjust appropriately. But letting an AI system or anyone else direct its own goals is not inherently harmful by itself, even if its goals are not fully aligned with many people, as long as we maintain a balance of power that favors people overall. The

main premise of the balance of power theory is that competing interests are natural, and the only way to ensure adherence to human values is by giving people more power.

This philosophy of personal empowerment, invention, and balance of power has several policy implications:

• **At Meta, we will focus our efforts on delivering personal superintelligence to billions of people and small businesses** to help everyone achieve their goals and invent the things they want to see in the world. We will do this by training leading models and then maximizing availability of this technology -- including making it easy to use, offering it for free or as affordably as possible, building sufficient compute for all people to use it, and distributing it widely. We will build a fully private mode where even Meta cannot see or grant access to your information. We will focus heavily on alignment -- defined as helping each person achieve their goals, not ours.

• **Open source is a positive and important force for empowering people** and preventing centralization that is detrimental for both safety and the economy. Meta continues to be strongly supportive of open source, including open source AI models. The current open source ecosystem is strong, and we think it would be a mistake to restrict it. Now that Meta Superintelligence Labs are up and running, we will resume releasing some open source models soon.

• **Independent governance and oversight is important** given the high stakes nature of decisions around superintelligence. I do not think it is in my, Meta's, or the world's best interests for me or anyone else to be a sole decision-maker on how superintelligence is deployed. Therefore, Meta is implementing a governance structure that gives our independent board of directors the power to approve the safety criteria for releasing models and reviewing whether each model release adheres to the criteria. While Meta is a founder-controlled company, the CEOs of all frontier labs currently have extensive authority over model releases, so we encourage others to implement structures similar to this as well. I think an industry-wide version of this process would be helpful for industry governance as well.

• **Government policy is necessary to ensure a positive future** and mitigating the risks discussed above. I have outlined several policy recommendations throughout this piece. Overall, we should encourage close cooperation between frontier labs and the government. It is important for national security that the government has the technical resources to harden critical systems and make effective use of advanced models. But it is also important for national security that the American AI industry continues to lead. We should accelerate the ability to build infrastructure -- whether that's energy, data centers, or silicon -- and build trust and alignment with communities across the country. We should be careful to not slow down American competitiveness. We should be skeptical of any proposals that lead to centralizing superintelligence, and we should favor policies that promote a healthy balance of power that favors people.

This is an incredible moment to live through. Developing superintelligence will be the most profound technological advance we will see in our lifetimes.

As we get closer to this, it is increasingly important to have a clear philosophy and values for how superintelligence will benefit humanity. Meta is committed to building with the principles of individual empowerment as the source of prosperity, invention as AI's purpose, and a balance of power favoring people as the foundation for addressing safety risks.

If these values lead the way, then I am optimistic that the coming decades will be some of the most amazing in history. The arc of human civilization has bent towards putting more power in people's hands to live and shape the world in the ways we believe are best. Superintelligence holds the promise of giving everyone that power, and building a positive future for everyone.

– Mark

August 10, 2026