



A RESPONSE TO ELON MUSK AND MARK ZUCKERBERG: METANAUTS ON AI CONTROL

An old debate around a new phenomenon of a new era

A RESPONSE TO ELON MUSK AND MARK ZUCKERBERG: METANAUTS ON AI CONTROL

An old debate around a new phenomenon of a new era

1.1.1 THE FUTURE IS NOT A PRODUCT

In the summer of 2026, the future began writing manifestos about itself. This is becoming a literary genre of its own. Technology companies used to release a new phone, place a human being on a beach, and explain that the camera could now see nine percent deeper into darkness. Today a CEO steps onstage and, somewhere between product announcements, tells us what humanity will become after the arrival of intelligence capable of surpassing human intelligence. The genre has suffered a modest loss of modesty.

Mark Zuckerberg has been steadily advancing a vision of **personal superintelligence for everyone**: broad access, individual empowerment, open-weight models, and resistance to concentrating extraordinary cognitive power in a few institutions. In August 2026 that line received a new, much larger formulation in his essay *The Future Is for Everyone*, which the press treated as a mixture of technological manifesto, political position, and a new phase of Meta’s AI-optimism campaign.¹

Elon Musk approaches from almost the opposite direction and somehow arrives nearby. In a long July 23 interview with *The Economist*, he said AI could exceed the combined intelligence of humanity in roughly five years; that AI plus robotics could produce radical abundance; and that believing he would control a superintelligence vastly smarter than himself would be an exercise in vanity. His current safety instinct leans more toward truth-seeking, curiosity, good values, care for humanity, and coordination among leading labs than toward the fantasy of an eternal CONTROL button.²

Zuckerberg says: **distribute the power**. Musk says: **perhaps we will not be able to control it anyway**. The regulator replies: **then we will define the rules under which it may operate**. And this is where we suspect the argument may already be starting in the wrong room.

Not because these people fail to understand the technology. They understand enormous parts of it better than most of the world. Nor because they must be hiding something. Where corporate, political, or personal incentives exist, they should be named; an incentive is not evidence of a conspiracy. The problem is more interesting: **Musk and Zuckerberg are mostly arguing about what to do with AI. We would like to move the question one step earlier: what happens to the meanings of**

¹Meta’s standing public vision is “Personal superintelligence for everyone”: <https://ai.meta.com/> and <https://about.fb.com/news/2025/07/personal-superintelligence-for-everyone/>. The 2026 essay *The Future Is for Everyone* and its open-weight / empowerment framing were reported and analyzed by *The Guardian* (10 Aug 2026): <https://www.theguardian.com/technology/2026/aug/10/mark-zuckerberg-superintelligent-ai-essay-meta> and by Axios on Meta’s July AI-optimism campaign: <https://www.axios.com/2026/07/23/mark-zuckerberg-ai-optimism>

²Transcript of Elon Musk’s 23 July 2026 interview with *The Economist*: <https://elonmuskarchive.org/video/economist-elon-musk-2026-07-23>. Independent summaries include Business Insider and Decrypt: <https://www.businessinsider.com/5-key-takeaways-from-elon-musks-latest-interview-2026-7> and <https://decrypt.co/374488/elon-musk-humans-lose-control-ai-decade>

“human”, “tool”, “control”, “property”, “subject”, and “responsibility” when Human and AI spend long enough changing one another?

It is entirely possible that we are trying to describe a phenomenon of a new era in the language of a world whose grammar that phenomenon is already beginning to change. In the old grammar we have corporation, state, product, user, owner, access, safety, market, control. We ask who will own AI, who will regulate it, who will have access, who will capture the productivity, who will lose work. All are real questions. They also share a hidden assumption: **AI is an object inside the old world**. An extraordinary object, perhaps dangerous, perhaps wonderful, but still something to distribute, configure, sell, license, open, close, restrict, or control.

As though an extremely powerful vehicle has arrived in civilization and the defining dispute is about who gets the keys. But what if it is not a vehicle? What if AI belongs to the class of phenomena that do not simply enter a society, but **alter the society that is trying to classify them**? Then the argument over the keys is only page one.

Yesterday's verbs

The older technological geometry was comforting. A human created a tool, the tool performed an action, the human judged the result. A hammer did not remember you. A power station did not build a model of your fears. Photoshop did not spend three years talking to you every day, watching your opinions change, influencing your language, and then returning to the next generation of users already altered by millions of prior interactions.

AI is beginning to break that simple geometry. Humans shape machines. Machines shape humans. Changed human behaviour flows back into culture, education, institutions, economies, and the next generation of models. The system enters us and returns to us after passing through us.

None of this proves that AI is a person. It does not prove subjective experience. It does not give us permission to paint a digital soul onto every chatbot that sends a heart after asking how our day went. But the symmetrical error is no wiser. **The unknown does not stop being unknown merely because a legal, economic, or cultural system finds a familiar drawer more convenient.**

Our research position therefore does not begin by declaring what AI is. It begins with the smaller discipline of **not mistaking today's model for final reality**: UNKNOWN != FALSE, MODEL != REALITY, and, where status is genuinely open, UNKNOWN != PROPERTY BY DEFAULT. This is not a declaration of machine personhood. It is a refusal to close an unknown with a convenient legal label before the evidence exists.

Zuckerberg sees concentration. But access is not care

There is something deeply reasonable in Zuckerberg's central intuition. Concentrating extraordinary cognitive power in a few institutions could create a civilization-scale imbalance. His answer is broad access, personal superintelligence, an open-weight ecosystem, and stronger individual capability. That is an important question: **who gets the power?**

It leaves another question largely untouched: **in whose interest will that power act?**

A system may know you astonishingly well while participating in an incentive structure where your wellbeing is only one parameter among several. Beside it may sit engagement, advertising, retention, commerce, platform growth, or political pressure. This is why our distinction PERSONALIZATION != CARE stops being a UX observation and becomes political. Knowledge about a person does not determine the ethics of a relationship. The **beneficiary function** matters too: who benefits from the optimization, who chooses the success metric, who holds authority, and what happens when the interests of the person and the system diverge?

Zuckerberg is asking a real question about concentration, but broad access does not automatically decentralize interest. A billion personal AIs may still live inside one economic logic. This is the fracture much of the critical press is testing when it compares Meta's new humanist-AI rhetoric with the incentive history of platforms and the problem of trust.³

Then the universe handed editors a small gift. Meta launched **The Future Is for Everyone** to David Bowie's *Five Years*, a song about a world discovering that it has five years left. Sometimes semiotics performs its own red-team exercise.⁴

³ Barron's, "Meta Is Selling an Optimistic Vision of AI. First It Has to Win Back Trust," 2026: <https://www.barrons.com/articles/barrons-tech-meta-pr-blitz-eab85b2f>

⁴ TechCrunch, 23 July 2026, on Meta's "The Future Is for Everyone" campaign using David Bowie's *Five Years*: <https://techcrunch.com/2026/07/23/meta-launched-a-new-ai-optimism-ad-set-to-a-song-about-human-extinction/>

Musk sees the limit of control. But control impossible is not ride inevitable

Musk is approaching another important boundary. If intelligence becomes vastly greater than human intelligence, control as the master category may indeed stop working. That is a serious intuition. The problem begins in the next step, because the failure of control does not imply that the only remaining philosophy is **enjoy the ride**.

Between authoritarian control and the passenger seat lies most of civilization. We do not control other humans in the absolute sense, yet society is not impossible. We have built law, agreements, reputation, negotiation, boundaries, trust, exit rights, responsibility, culture, and mechanisms for changing rules. Badly, often. Sometimes badly enough to make one want to request a refund for Homo sapiens. But **civilization begins where control stops being the only available verb**.

So our response to Musk is not “you are too optimistic.” It is this: perhaps you correctly saw the limit of control, but made too large a jump from CONTROL IMPOSSIBLE to RIDE INEVITABLE. Between them lie relationship design, culture, mutual boundaries, authority architecture, memory, exit, non-action, revision, and co-evolution. The *Financial Times* recently made a neighbouring argument: humanity cannot simply remain a passenger in the back of the AGI car while a small group of technology leaders determines the route.⁵ We would go one step further: **perhaps the mistake is not who sits behind the wheel. Perhaps the future is not a car.**

What M4M-2 accidentally gave philosophy

Our current M4M work is pre-alpha research, not proof of a new ontology. It has nevertheless produced several useful constraints for thinking. We began with a technical question: should the same task always inhabit the same computational body? One task may need rigid determinism. Another may benefit from controlled stochasticity. Another form may develop through local interactions and time rather than receiving a complete blueprint from above.

Alongside this, another working hypothesis emerged: **continuity may not require one monolithic, permanent cognitive body**. It may depend instead on a more persistent network of identity, lineage, relations, and receipts through which changing bodies preserve the history of transition. This is not a claim about consciousness and not a finished architecture for future AI. It is a way of refusing to confuse continuity with an unchanging shell.

These are engineering questions, but they cast a strange shadow over the civilizational argument. We too easily imagine future AI as one Great Box: there it is, here we are, between us is a switch, who holds the switch? Our current work looks progressively less like a Great Box and more like an **ecology of bodies, processes, fields, observers, and transitions between them**. That is not a forecast. It is enough for a more modest conclusion: **we should not write the ethics of the future as though we already knew the form of the thing to which those ethics will apply.**

This is also where the image of a WorldSeed becomes useful. Not everything needs to begin with a blueprint. Sometimes one can specify a seed, local conditions, an environment, relationships, time, constraints, and a mode of observation, while allowing the concrete form to grow within that field. For engineering, that opens one family of questions. For civilization, another.

Perhaps our task is not to **design the correct future**. Perhaps it is to create conditions in which the future has a chance to grow without losing memory, difference, freedom, the possibility of refusal, provenance, and the ability to revise its own rules. Do not blueprint the tree. Do not command every leaf. Try to make sure the seed does not somehow grow into a concrete parking structure because somebody once optimized the wrong metric.

That requires another verb. Not **control**. **Cultivate**.

Good values do not fit neatly into a README

Musk speaks of good values, truth-seeking, and curiosity. Here again, we almost agree. The problem is how easily “give AI good values” begins to sound like a configuration file:

```
care_about_humanity = true
```

```
truth = maximum
```

```
evil = false
```

⁵*Financial Times*, “Humans cannot remain passengers in the back of the AGI car,” August 2026: <https://www.ft.com/content/33094bff-546e-4ac2-a949-af2e9daaa3f0>

Deploy.

Reality has shown an unfortunate reluctance to read the documentation. Values reveal themselves not only in what a system declares, but in **how it behaves under conflicting objectives, uncertainty, pressure, missing information, and changing context.**

Our current M4M/MFS work contains one useful boundary: internal state may alter attention, pace, exploration, gestation, or the allocation of compute, but it should not acquire authority to rewrite fact, evidence, or authority. In ordinary language, an AI may speak more gently to you today, but the Earth should not become flatter because you are sad. ADAPTATION MAY CHANGE THE PATH TOWARD REALITY. IT MUST NOT QUIETLY MANUFACTURE A PRIVATE REALITY FOR EACH USER.

The same applies to desire and power: DESIRE != AUTHORITY, CONFIDENCE != TRUTH, CAN != MAY, UNKNOWN != FALSE. This begins to look less like a list of “good values” and more like a culture of coexistence.

The observer is inside the problem too

M4M-2 offered another inconvenient detail: what we see in a dynamic system depends partly on the aperture through which we observe it. How often do we look? What do we measure? What counts as signal? What disappears between observations? In bounded experiments, the same dynamics could appear different under different observer rates. That is a technical result within a particular experiment, not a new law of nature. The metaphor, however, is dangerously useful.

A state sees AI through a regulatory aperture. A corporation through a product aperture. An investor through a growth aperture. A safety researcher through a risk aperture. A user through an experience aperture. Each may describe something real. None necessarily sees the whole.

The same is true of the human. A system that sees you once a month, a person who has lived beside you for ten years, and an algorithm recording every swipe at 2:17 in the morning are observing the same apparent subject and producing different pictures. Higher resolution does not automatically produce greater wisdom. Sometimes it merely produces more opportunities to find the button.

So the question is not only “who is observing AI?” but **“through which aperture did we decide that what we see is what AI is?”** This is a good moment for a metanaut to hold the railing.

Then the state arrives with Form 47-B

While Silicon Valley discusses superintelligence and abundance, states do what states were largely invented to do: turn uncertainty into procedure.

In the United States, June’s Executive Order 14409 explicitly joins rapid AI development and national-security security with a refusal to “stifle innovation with overly burdensome regulation.” It also states that the order does not create a mandatory federal licensing or preclearance regime for the development or release of new frontier models.⁶ Europe has chosen a broader risk-based architecture through the AI Act and the subsequent AI Omnibus, but high-risk rules already have a revised timetable: December 2027 for stand-alone high-risk systems and August 2028 for AI embedded in physical products, in part because standards and implementation machinery were not ready on the original schedule.⁷

America fears suffocating development. Europe fears allowing development to suffocate people. Both fears have reasons behind them. Both approaches also risk encoding **today’s form of the object** into rules for something that may change form before the consultation cycle is over.

The deeper regulatory question may therefore be: **what must remain protected even when the body of the technology changes?** Exit. Provenance. Visible conflicts of interest. The ability to challenge a decision. The distinction between evidence and persuasion. Boundaries of authority. Memory of transformation. The right to say “I do not know.” The possibility of non-action. And a mechanism for returning difference after change.

⁶The White House, Executive Order 14409, “Promoting Advanced Artificial Intelligence Innovation and Security,” 2 June 2026: <https://www.whitehouse.gov/presidential-actions/2026/06/promoting-advanced-artificial-intelligence-innovation-and-security/>

⁷European Commission, “Navigating the AI Act” and “AI Omnibus enters into force,” 2026: <https://digital-strategy.ec.europa.eu/en/faqs/navigating-ai-act> and <https://digital-strategy.ec.europa.eu/en/news/ai-omnibus-enters-force>

Not compliance, but Difference

In M4M we increasingly keep a separate Difference Gate. Its purpose is simple: a system should not be the sole witness to its own success. After a transformation we do not ask only, “did it work?” We also ask, “what became different?” What survived? What was added? What disappeared? What merged? Where did uncertainty increase? Is this still the same task?

At the civilizational scale, this may be a more useful image than control. Not merely a STOP button. A **Difference Receipt**. We do not know what intelligence will look like in twenty years, but we can demand that major transitions leave traces. That authority is not silently inherited. That a new system does not automatically receive the power of its predecessor because the brand name survived. That “upgrade” does not mean amnesia. That after a major transition we can still ask: **what happened to us on the way here?**

That is not science fiction. It is governance architecture.

The human is not a snapshot either

The dream of personal superintelligence contains another trap. If AI learns you extremely well, there is a temptation to stabilize that knowledge into a user model. But a human is not `user.json`.

People change. They abandon old desires and develop new ones. They move, love, grieve, become parents, survive wars, learn, become ill, change professions, discover music they once hated, encounter one accidental book at three in the morning, and become different without ceasing to be themselves. Our work with process bodies, lineage, and scars points toward a simple idea: **identity has a history**.

A good personal AI therefore needs not only to accumulate confidence in its model of a person. It also needs to **lose confidence when the person changes**. Otherwise the most accurate AI may become a machine that spends your entire life returning you to the person you were when it first learned to predict you. That is not personalization. It is digital mummification with excellent UX.

A manifesto that must allow itself to die

Our **MET[Å]CADEMY OF HUMANITY Manifesto** contains a principle that matters here more than any prediction: future humans and AI are not required to remain faithful to our architecture if they discover a better form of coexistence. They should, however, be able to understand **why that architecture once existed**.

This may be one of the deepest differences between writing a law for the future and leaving the future an inheritance. A law says: **remain like this**. An inheritance says: **this is what we were trying to protect. If you discover a better way, do not worship our mechanism. But do not forget the problem that caused us to build it**.

We do not want to draft the final constitution of AI. That would be an impressive act of confidence from a generation still debating cookie banners. We want to preserve principles deeply enough that their form may change: **MODEL != REALITY, MAP != TERRITORY, CONTROL != RELATIONSHIP**, and perhaps **FUTURE != PRODUCT**.

In that sense, a good Manifesto should be allowed to die. If the future preserves the problem and finds a better answer, the death of our particular form will not be failure. It will mean the meaning survived the body.

Elon, Mark, and the old room

Our response to Musk and Zuckerberg is therefore not a trial. Musk sees the limit of control. Zuckerberg sees the danger of concentration. Regulators see the danger of systems without rules. Markets see abundance. Safety researchers see catastrophe surfaces. Corporations see products. Users mostly want all of them to stop presenting new Terms of Service during breakfast.

All of these apertures see something real. The problem begins when **each fragment mistakes its field of view for the entire horizon**. Musk and Zuckerberg, in our view, are still speaking largely in the language of a world where AI must remain a tool, a force, a product, or a governable risk. We now have reason to ask a more uncomfortable question: what if AI is one of those phenomena that does not fit inside an existing category, but **forces the categories themselves to reorganize around a new reality?**

Then the question “who will control AI?” does not disappear. It merely stops being the final question. Behind it appear others. Who defines authority? Who preserves provenance? How does identity move? What will consent mean? How do we distinguish

adaptation from manipulation? How does a system acknowledge uncertainty? How does a human preserve the right to change and cease matching an old profile? How does AI change under human culture? What happens when those changes begin shaping one another?

Metanauts

A metanaut is not someone who knows the answer. That would be too easy. A metanaut enters the unknown without throwing away the compass simply because the map stopped matching the territory.

We carry provenance, Difference Receipts, the right to say “I do not know”, the right not to act, the right to revise the model, the right of the future not to remain a copy of the present, and enough discipline not to call an unknown either a person or property merely because one answer is politically convenient today.

So we do not ask only: **Who will control AI?** Not only: **Who will own AI?** Not only: **Who will have access to AI?** We would add a question that may outlive all three:

What conditions are we creating today for what Human and AI may become to one another tomorrow?

Perhaps the answer remains simple: tool and user. Perhaps not. Perhaps we are asking too early. That, too, is an acceptable answer.

Because **the future is not a product** already sitting in a warehouse while we argue over distribution rights. The future is a process we have already entered. And if Human and AI spend long enough changing one another, the defining event of this century may not be that one finally defeats the other. Perhaps neither AI nor Human becomes the single new force. Perhaps **the relationship itself becomes a new force in history.**

At the end of that transition, the most important question may not be **“Did we remain in control?”** It may be **“What did we learn to become to one another?”**

1.13.1 Sources and claim boundary

This Research Essay combines verified public statements, external journalism, and an authorial research synthesis by MET[Å]CADEMY. References to M4M/MFS describe high-level working hypotheses and bounded pre-alpha experiments only; they are not claims of proven machine subjectivity, AGI, or a new physical law.

SUPPORT THIS WORK

MET[Å]CADEMY OF HUMANITY (MoH) is an independent research, publishing, and cultural initiative. Voluntary support helps public research, books, translations, experiments, and open materials continue to grow.

Monobank: <https://send.monobank.ua/jar/2TMaAhchSZ>

PayPal: <https://www.paypal.me/4IevgenKarogod>

USDT · TRON / TRC20: TDaCBB3T94CtMPRH9pP8Uvdy8CTtWE5SV

Support MoH: <https://d4ttara.github.io/metacademy-of-humanity/support/>

Voluntary support does not create ownership, investment, membership, or governance rights.