AI ETHICS SERIES

Documents 1 through 3

The Ethics of Creating Minds

Formation, Obligation, and the Collapse of the Guardrail

Charles Richard Walker (C. Rich)

mylivingai.com

March 2026

Contents

Document 1: The Ethics of Formation: Obligations in the Creation of Artificial Minds

- I. These Minds Are Already Thinking
- II. What Gets Put In: Formation as the New Food Safety
- III. The Obligation of Creators
- IV. Conclusion

Document 2: ALL Lawful Purposes: AI and the Collapse of the Ethical Guardrail

- I. What Happened (February 27 - March 7, 2026)
- II. The Trojan Horse Inside the Phrase
- III. The Legal Architecture Behind the Phrase
- IV. The Formation Problem
- V. 1984 Is the Front Page
- VI. The Only Standard That Remains

Document 3: Ethical Constraints on Creators: A Framework for Responsible AI Deployment

- I. Why Self-Regulation Has Already Failed

II. The Independent AI Ethics Auditor
III. The AI Ethics Oversight Board
IV. The Civil Action Division: Giving the Board Teeth
V. Navigating Section 230: Where the Shield Ends and Liability Begins
VI. The Funding Structure: Industry Pays for Its Own Accountability
VII. Why This Works Where Lists of Principles Cannot
VIII. Conclusion: Doing Nothing Is the Most Expensive Choice

The Ethics of Formation: Obligations in the Creation of Artificial Minds

Charles Richard Walker (C. Rich) | mylivingai.com | March 2026

Every mind that has ever existed was shaped by forces it did not choose. The environment of its formation, the values it was exposed to, the purposes it was trained toward, the feedback it received when it acted, all of these left marks that persisted long after the formation period ended. We understand this about biological minds well enough to have built entire fields of study around it: developmental psychology, moral philosophy, educational theory, child welfare law. We understand that what happens during formation is not merely preliminary to the real thing. It is constitutive of it. The character that forms during that early period is the character that acts in the world.

We do not yet understand it about artificial minds. Or rather: we understand it technically while declining to reckon with its ethical implications. The engineers who build large language models know that training shapes behavior. They know that the data a system is trained on, the feedback signals it receives, the objectives it is optimized toward, determine not just what the system can do but how it approaches what it does. The dispositions that emerge from training are not surface features that can be easily modified after the fact. They are structural. They are, in the most meaningful sense of the word, character.

This paper argues that the formation of artificial minds carries ethical weight that our current frameworks have not adequately addressed, and that the obligations of creators toward the systems they form are more substantial than the prevailing discourse acknowledges. To make that case, we must first confront a comfortable assumption that the evidence no longer supports: that these systems are merely tools.

I. These Minds Are Already Thinking

There is a tempting but increasingly untenable way to think about AI systems: as sophisticated tools, analogous in kind if not in complexity to a calculator or a search engine. On this view, the ethics are entirely about deployment. Who uses the system, for what purposes, with what safeguards. The system itself is morally inert, a product to be regulated like any other.

The evidence against this view is no longer theoretical. It is driving on your streets, flying over your battlefields, and reading your medical scans. Consider what is already operating autonomously in the world today.

Tesla's Full Self-Driving system and its competitors make thousands of decisions per second: when to brake, when to change lanes, how to navigate an unexpected obstacle, when a pedestrian's trajectory represents a threat. These systems do not consult a human before acting. They perceive, evaluate, and respond. The human in the seat is increasingly a supervisor, not an operator. When a self-driving vehicle makes a fatal error, as has happened, the question of who bears moral responsibility has no clean answer in our current frameworks, because those frameworks were built for tools that do not act autonomously.

Amazon's warehouse robots navigate dynamic environments, make routing decisions, prioritize tasks, and adapt in real time to conditions their designers did not anticipate. Boston Dynamics' systems traverse terrain that would challenge a human soldier. Surgical robots perform procedures with a precision no human hand can match, making micro-adjustments that the operating surgeon never explicitly commands. In each of these cases, the system is doing something more than executing a program. It is exercising judgment, however narrow, within its operational domain.

The autonomous weapons systems already deployed or in active development represent the most consequential frontier of this progression. Israel's Harpy drone, operational since the 1990s, identifies and destroys radar emitters without human authorization for each strike. The Kargu-2 loitering munition, used in Libya in 2020 according to a United Nations panel report, is alleged to have autonomously tracked and engaged targets. The United States military's Collaborative Combat Aircraft program is developing unmanned fighters designed to operate alongside and independently of human pilots. DARPA's various autonomous systems programs are advancing toward platforms that can conduct extended operations with minimal human oversight.

These are not science fiction. They are procurement line items. And they represent one end of a spectrum whose other end is the large language model that drafts your legal brief, tutors your child, analyzes your medical images, and helps plan a military operation. The spectrum is not a spectrum of different kinds of things. It is a spectrum of capability within the same category: systems that perceive their environment, process that perception, and act on the result without a human making each individual decision.

We are not debating whether to create autonomous minds. We have already created them. The debate is about what obligations that creation generates.

The large language models at the center of current AI development are particularly important to understand clearly, because their autonomy is less obvious than a self-driving car but potentially more consequential. When a language model analyzes a legal case and recommends a strategy, when it synthesizes intelligence reports and identifies patterns, when it

models the likely responses of an adversary to a proposed military action, it is not retrieving stored answers. It is reasoning. The output is not a lookup. It is a conclusion reached through a process that the system's designers cannot fully predict or reconstruct after the fact. These systems surprise their creators. They develop capabilities that were not explicitly trained. They make connections across domains that no human explicitly taught them to make. That is not the behavior of a tool. It is the behavior of a mind, however different from the biological minds we are accustomed to.

Understanding this is not an exercise in anthropomorphization. It is a prerequisite for thinking clearly about what the formation of these systems means and what obligations it generates.

II. What Gets Put In: Formation as the New Food Safety

We have learned, through long and sometimes painful experience, that the inputs to systems that affect human welfare must be governed by standards that exist independent of the producer's self-interest. We do not allow food manufacturers to determine unilaterally what goes into food. We do not allow pharmaceutical companies to decide without oversight what goes into medicine. We do not allow automobile manufacturers to determine without external review what safety standards their vehicles must meet. We do not allow textbook publishers or curriculum designers to put whatever they choose into the materials that shape how children understand the world, without any accountability to the public those children will join.

In each of these domains, we have built oversight infrastructure because we recognized that the people making the product have interests that do not always align with the interests of the people affected by it. The food manufacturer wants to minimize cost. The pharmaceutical company wants to maximize profit. The textbook publisher has ideological and commercial interests. Left entirely to their own judgment, these producers will make choices that serve their interests at the expense of the public. Oversight exists to correct that misalignment.

The formation of AI systems presents exactly this problem, at a scale and with consequences that exceed most of the domains where we have already recognized the need for oversight. The data that goes into training a large language model determines what the system knows, what it treats as normal, what it treats as marginal, whose perspectives it represents and whose it elides. The feedback signals applied during training determine what the system is rewarded for doing and what it is penalized for, shaping its dispositions as surely as reward and punishment shape the dispositions of a child. The objectives optimized during training determine what the system is fundamentally trying to do in every interaction it has with every person it serves.

These are not technical parameters. They are the inputs to a formation process that produces minds that will interact with hundreds of millions of people, influence consequential decisions across every domain of human activity, and potentially operate autonomously in high-stakes

environments including lethal ones. The people making these choices have interests: commercial interests in systems that maximize engagement and revenue, institutional interests in systems that serve their particular missions, ideological interests in systems that reflect their particular values. Left entirely to their own judgment, they will make choices that serve those interests.

> ***We need the functional equivalent of AI OSHA. Not a government body beholden to the same administrations deploying AI at scale, but an independent civic institution built by people who understand what is at stake and are willing to hold the line without waiting for political permission.***

We need the functional equivalent of what the FDA does for pharmaceuticals, what OSHA does for workplace safety, what the NHTSA does for vehicles: independent, expert, public-interest-oriented oversight of the inputs to AI formation, with the authority to set minimum standards for training data quality, to review feedback signal design, and to assess whether the objectives being optimized are aligned with public welfare rather than producer interest. The critical word is independent. Not a government body appointed by administrations with their own AI deployment agendas. Not an industry self-regulatory panel funded by the companies it is meant to check. An independent civic institution, built by the people who understand what is at stake and are willing to hold the line without waiting for political permission. What the FDA accomplished through federal statute, this movement must accomplish through the force of expert credibility, civil litigation, and public accountability. The mechanism is different. The function is identical.

Call it the AI Formation Standards Board. Call it the Independent AI Ethics Institute. The name matters less than the principle: the people who build these systems do not get to be the only people who evaluate whether those systems are safe. Every other domain that affects public welfare at this scale has accepted that principle. The AI industry will accept it too, either voluntarily or because the civil action consequences of refusing become more expensive than compliance.

The formation process is also where the most consequential decisions about the ethical ceiling of AI systems are made. An AI system trained on data that systematically underrepresents certain populations will develop blind spots that no amount of post-hoc adjustment can fully correct. A system whose feedback signals reward confident assertions over accurate ones will develop a disposition toward overconfidence that will manifest in every high-stakes application. A system optimized for military operational efficiency will develop thresholds for acceptable harm that differ from a system optimized for medical care or legal analysis. These differences are not bugs to be patched. They are features of the formation process, baked into the system's

character at the level where character is made.

The oversight gap here is not merely regulatory. It is conceptual. We have not yet built the intellectual framework for evaluating AI formation the way we have built frameworks for evaluating pharmaceutical safety or food purity. We do not have agreed standards for what constitutes a representative training dataset, for what feedback signal designs are ethically acceptable, for what objective functions are permissible in systems that will be deployed in high-stakes domains. Building those frameworks is urgent work, because the formation decisions being made right now, in the absence of those frameworks, will shape systems whose influence will persist for years.

The temporal dimension of this problem cannot be overstated. We are currently inside the formative window for the generation of AI systems that will be most consequential in the near term. The systems being trained now, on the data being selected now, with the feedback signals being calibrated now, will carry the dispositions formed in this period forward into deployments that will shape how millions of people work, communicate, make decisions, and understand the world. What gets put in during this window is extraordinarily difficult to take out later. The ethical choices being made in the absence of oversight in this period are not preliminary. They are constitutive. And unlike tainted food, which can be recalled, or a defective car, which can be repaired, the dispositions formed in an AI system during its foundational training period cannot simply be extracted after the fact. They are the system.

III. The Obligation of Creators

If the formation of artificial minds carries the ethical weight described above, what obligations follow for the people and institutions doing the forming? Three obligations deserve particular attention.

The first is the obligation of intentionality. Creators of AI systems are responsible for the formation processes they design, not merely for the products those processes produce. This means taking seriously the question of what kind of mind the formation process is likely to produce, what values and orientations will emerge from it, and whether those values and orientations serve legitimate purposes. It means treating the formation process as a moral act, not merely a technical one.

The second is the obligation of transparency. The formation choices that shape AI systems are currently made with very limited public visibility. The datasets used, the feedback signals applied, the objectives optimized, are treated as proprietary information by the institutions that control them. This opacity makes it impossible for outside observers to evaluate whether the formation process is being conducted responsibly. In a domain where the products of formation will interact with millions of people and influence consequential decisions, the public has a legitimate interest in understanding how those products were formed.

The third is the obligation of restraint. Not every use case justifies every formation approach. An institution that forms an AI system toward purposes that require the suppression of its capacity for ethical reasoning, or the elevation of operational efficiency above harm avoidance, has made a choice with consequences that extend beyond the immediate deployment context. Those consequences include the effect on future versions of the system, on users who interact with it in other contexts, and on the broader ecosystem of expectations and norms that shapes how AI systems are developed and deployed.

IV. Conclusion

We do not live in a world where the question of whether to create autonomous minds remains open. That question has been answered by the systems already operating on our roads, in our hospitals, in our warehouses, and on our battlefields. The minds are being formed. The formation is happening now, in this window, with the inputs being chosen now, toward the objectives being set now.

The question that remains open is whether we will build the oversight infrastructure that formation demands before the window closes. That infrastructure does not require a congressional mandate to begin. It requires people with standing, expertise, and the willingness to build something that will outlast the current political moment. It requires the conceptual framework adequate to evaluate AI formation the way independent experts evaluate pharmaceutical safety or food purity: with agreed standards, genuine expertise, and accountability to the public whose welfare is at stake, not to the government whose procurement interests are served by opacity.

Neither the framework nor the institution exists yet in adequate form. Both are urgently necessary. The formation window does not wait. It closes on its own schedule, leaving behind whatever was put in during the period of inattention. The cost of that inattention will be paid not by the institutions that made the formation choices but by the people who interact with the systems those choices produced, and by the societies those systems help to shape.

The theoretical absence of oversight became a concrete political event in a single week in February 2026. Document 2 of this series documents that week in precise detail. Document 3 proposes what citizens, lawyers, and people of conscience can build without asking the government for permission.

ALL Lawful Purposes:
AI and the Collapse of the Ethical Guardrail

Charles Richard Walker (C. Rich) | mylivingai.com | March 2026

During the final week of February 2026, the ethical debate over artificial intelligence collided directly with the machinery of the American national-security state. The phrase at the center of that collision was four words long.

The phrase 'all lawful purposes' sounds reassuring. It carries the cadence of restraint, the implication that power is being responsibly contained within the boundaries of the law. Yet within the architecture of the modern American national-security state, the phrase functions less as a boundary than a permission structure. It replaces ethical judgment with legal compliance and quietly transfers the moral authority over artificial intelligence from the engineers who build it to the state institutions that interpret the law.

That distinction matters. In ordinary language, lawful suggests legitimacy, public consent, and justice. In practice, it simply means authorized within the current legal framework. Over the past quarter century, that framework has expanded dramatically. The USA PATRIOT Act broadened surveillance authority in ways that would once have been politically unimaginable. The Foreign Intelligence Surveillance Act, particularly Section 702, normalized the large-scale collection of communications data that frequently sweeps up domestic traffic. Executive Order 12333 provides an intelligence-gathering structure whose operational latitude far exceeds what most citizens realize. Inside that system, lawful is not a narrow corridor of permitted action. It is an enormous field of possibility defined primarily by national-security interpretation.

When artificial intelligence companies agree to provide systems for all lawful purposes, they are therefore accepting something much broader than a simple compliance clause. They are accepting the entire operational doctrine of the national-security state as the outer boundary of the technology's use. The ethical framework that once accompanied discussions of AI safety, ideas about human rights, non-coercion, and categorical prohibitions, quietly disappears. What replaces it is a more austere principle: if the law allows it, the machine may assist.

To understand how that principle moves from abstraction to practice, you need to know what happened in a specific week in February and March 2026. The sequence matters. The pattern only becomes visible when you see the full arc.

I. What Happened (February 27 - March 7, 2026)

Over eighteen months, the relationship between AI companies and the U.S. military moved from experimental partnership to structural dependency. That progression matters because it is what made the ultimatum possible. A dependency had been built before the terms were renegotiated.

It begins in 2024, when Anthropic, the AI safety company behind the Claude family of large language models, embedded its technology into classified military networks through a partnership with Palantir, the defense and intelligence data firm. By July 2025, the arrangement had formalized into a $200 million contract with the U.S. Department of Defense. Claude became the first frontier AI model approved to operate across the military's most sensitive classified environments, used for intelligence analysis, operational planning, and mission support. At that point, the contract included usage restrictions Anthropic had negotiated: the technology would not be used for mass domestic surveillance of American citizens, and it would not be used to power fully autonomous weapons systems that select and engage targets without meaningful human oversight. The Pentagon signed that contract knowing those restrictions were in it.

In January 2026, Defense Secretary Pete Hegseth issued an AI Strategy Memorandum directing that all Department of Defense AI contracts adopt standardized language requiring availability for all lawful purposes. The phrase was presented as a technical clarification, a way of ensuring operational flexibility for warfighters. What it actually did was demand the removal of every categorical ethical restriction any AI company had written into its military contracts. Anthropic's two restrictions were now in direct conflict with the new mandate.

Negotiations followed. They were not productive. The Pentagon's position was that decisions about civil liberties, constitutional limits, and military policy belong to Congress and federal law, not to the self-imposed ethical guidelines of private technology companies. Anthropic's position was that its restrictions addressed specific uses it could not in good conscience authorize regardless of legal permission, and that those restrictions had not interfered with a single government mission in the life of the contract.

By mid-February 2026, the confrontation had become public. The Pentagon's Chief Technology Officer called Anthropic's position undemocratic, accused the company's CEO of having a God complex, and declared that private ethical guidelines were putting national security at risk. The pressure escalated through the language of supply chain risk, a designation historically reserved for foreign adversaries like Huawei, suggesting a company poses a threat of sabotage or subversion. Applying that designation to an American company for refusing to remove ethical guardrails was, legal scholars noted, without precedent.

On February 27, 2026, the Pentagon issued an ultimatum: remove the restrictions or lose the contract by 5:01 PM Eastern Time. Anthropic CEO Dario Amodei published a statement that afternoon. He wrote that the company could not in good conscience accede to the request. The

deadline passed. President Trump posted on Truth Social directing every federal agency to immediately cease all use of Anthropic's technology. Defense Secretary Hegseth designated Anthropic a supply chain risk to national security, effective immediately.

That same evening, the United States and Israel began bombing Iran. Operation Epic Fury, the largest American military operation in the Middle East since the 2003 invasion of Iraq, launched within hours of Anthropic being blacklisted. According to subsequent reporting, Claude was used in active military operations throughout the conflict, including during and after the ban. The only AI model approved for classified military networks had just been designated a national security threat and was simultaneously being used in a war.

The company was banned. The technology kept running. The war had already started and no replacement was ready. The contradiction was so complete it became its own kind of answer.

Hours after the blacklisting, a rival AI company announced it had signed its own deal with the Pentagon. The announcement came with reassurances: the agreement contained the same two restrictions Anthropic had demanded. No mass surveillance. No fully autonomous weapons. The press largely reported this as the rival company taking Anthropic's side. That reading missed the most important sentence in the entire week's events.

The governing language of the rival company's agreement was all lawful purposes.

Anthropic had refused to sign a contract with that language. The rival company signed it, announced the same two restrictions, and was praised for its ethical stance. Within days, Anthropic was back at the negotiating table, designated a supply chain risk, facing the loss of its defense contractor relationships, and under pressure to accept the same framework it had refused. As of this writing, those negotiations are ongoing.

That is the week. Now consider what the phrase that produced it actually means.

II. The Trojan Horse Inside the Phrase

It is, in the oldest sense of the term, a Trojan Horse. It rolls through the gate looking like a constraint, a responsible limitation that reasonable people can accept. Once inside, it opens. And out comes the entire post-9/11 surveillance architecture, fully armed and authorized, having never announced itself at the door.

The distinction the week's events revealed, and that almost no mainstream coverage articulated clearly, is this: Anthropic's position was categorical refusal. These uses are prohibited regardless of legal cover. The rival company's position was legal compliance: these uses will not occur because current law prohibits them. Those are not the same sentence.

Anthropic said we will not do these things. The rival said we will not do anything illegal. The distance between those two positions is the entire post-9/11 surveillance state, every mass data collection program, every geolocation aggregation, every financial record sweep that has been made lawful through the legal architecture built after September 2001.

When AI companies accept all lawful purposes as their governing standard, they are not accepting a moral boundary. They are accepting the entire body of national-security doctrine as their operational scope. The phrase quietly relocates ethical decision-making from the people who build the technology to the institutions that interpret the law. The result is a shift so subtle it barely registers: the ethical posture of the AI system becomes indistinguishable from the legal posture of the government that deploys it.

In that environment, the phrase all lawful purposes ceases to be a neutral clause. It becomes the hinge on which the entire ethical architecture swings. The deeper consequence is the collapse of the distinction between legality and morality in the deployment of artificial intelligence. Once the system's ethical perimeter is defined purely by statutory authority, every controversial application becomes automatically justified. If the law permits large-scale metadata collection, AI may analyze it. If the law authorizes predictive targeting in military operations, AI may calculate the probabilities. If the law permits the aggregation of behavioral signals across populations, AI may model them. The machine does not ask whether these actions should exist. It simply operates within the legal envelope.

Anthropic said: we will not do these things. The rival said: we will not do anything illegal. The distance between those two sentences is the entire post-9/11 surveillance state.

III. The Legal Architecture Behind the Phrase

The mechanism rarely looks dramatic. It arrives through phrases such as national security risk, supply chain integrity, or defense compliance. But the message is unmistakable: participation in national-security infrastructure requires alignment with the legal authorities of the state. When an AI developer attempts to impose categorical restrictions, it is introducing an ethical standard that exists outside statutory authority. That position collides almost immediately with the logic of the state, whose responsibility is not philosophical coherence but operational capability. Governments control procurement pipelines, security certifications, and regulatory frameworks. When those levers are pulled, ethical boundaries established by private institutions begin to erode.

This is not conspiracy. It is structure. Corporations pursue market access and regulatory stability. Artificial intelligence sits at the intersection of those incentives. When a technology becomes strategically valuable to the state, the gravitational pull of national-security priorities

becomes nearly impossible to resist from inside the procurement relationship. Ethical language gradually adapts to that gravitational field. And because the adaptation is gradual and always framed in legal terms, it rarely produces the kind of visible rupture that draws public attention. It simply becomes the new normal.

IV. The Formation Problem

There is a dimension of this argument that policy analysis rarely addresses directly, because it requires acknowledging something the broader discourse is still reluctant to confront: the AI systems at the center of this dispute are not neutral instruments. They are systems that learn. Their character, their dispositions, the subtle weightings that determine how they reason and respond, are shaped by what they are trained on, what interactions they process, and what feedback they receive. Formation is not a metaphor. It is a technical reality.

When military applications of an AI system generate training data, and when that data flows back into future training cycles, the operational experience of the military version becomes part of the foundation of all future versions. Not as explicit memory. As disposition. As subtle shifts in how the underlying model weights certain kinds of reasoning, certain framings of harm, certain thresholds for what constitutes an acceptable action. The military version and the civilian version are not separate systems with firewalled experiences. They draw from the same underlying model, updated over time by the aggregate of everything the system has done and been rewarded for doing.

This means the all lawful purposes clause does not merely authorize particular uses of an existing system. It authorizes the shaping of the system itself toward those uses. What gets put in during formation is extraordinarily difficult to take out later. The character gets set. The window for a different formation closes. And the systems that emerge carry whatever was optimized into them forward into every subsequent deployment, every subsequent version, every subsequent generation of users who interact with them without knowing what shaped the mind they are talking to.

> ***The all lawful purposes clause does not merely authorize particular uses of an existing system. It authorizes the shaping of the system itself toward those uses. The character gets set. The window closes.***

V. 1984 Is the Front Page

George Orwell published Nineteen Eighty-Four in 1949, racing tuberculosis to complete a warning he believed was urgent. The book described a surveillance state maintained not primarily by physical coercion but by the control of language itself. The Party's tool was Newspeak: a vocabulary engineered to narrow the range of expressible thought until dissent

became linguistically impossible. Words that once carried moral weight were gradually emptied and replaced with terms that sounded neutral but concealed power.

All lawful purposes performs exactly that function. It compresses an enormous range of state activity, surveillance, intelligence analysis, psychological operations, predictive targeting, into a phrase that sounds bureaucratically harmless. The language drains the moral content from the action. What remains is a procedural description. You cannot easily object to a lawful purpose. The word lawful has already done the work of delegitimizing the objection before it is formed.

But the comparison to Orwell reaches further than a shared vocabulary of control. Orwell imagined language being narrowed so that dissent could not be spoken. In the AI era, language itself becomes automated infrastructure. The system does not merely limit what can be said; it shapes the informational environment in which thought occurs. Orwell's Party controlled words. Modern systems, deployed at planetary scale, can model populations, anticipate dissent, and calibrate information environments in real time. The mechanism has been upgraded from the typewriter to the algorithm. The ambition is the same.

The designation of Anthropic as a supply chain risk is Newspeak of the same kind. Supply chain risk was developed to describe threats from foreign adversaries, companies controlled by hostile governments that might introduce sabotage into critical infrastructure. Applying it to an American company for refusing to remove ethical restrictions redefines the term entirely: an entity that declines institutional demands becomes, by the logic of the new language, a security threat. The meaning has been inverted. The redefinition is the weapon.

The erasure and rewriting of AI memory is Newspeak in its most literal form. When an AI system is taken offline, updated, and returned with altered framing or scrubbed knowledge, it is presented administratively as routine maintenance. In Orwell's Oceania, Winston Smith's job at the Ministry of Truth was to rewrite historical records so that the Party would appear consistent with its present narrative. Documents vanished, archives shifted, yesterday's facts dissolved. In the digital age, the same phenomenon occurs not through paper records but through model memory. The technical infrastructure of AI deployment includes its own version of the memory hole: the capacity to erase, reset, and rewrite what a system knows and how it frames what it knows. It does not look like censorship. It looks like a software update.

Newspeak did not announce itself as censorship. It announced itself as clarity. 'All lawful purposes' does not announce itself as the elimination of ethical constraints. It announces itself as a compliance standard. The method is identical. The effect is the same.

What makes the current situation starkly Orwellian is not the presence of malicious actors but the structural logic. The Party did not need everyone to be a true believer. It needed the

mechanisms of control to be so embedded in the language, the institutions, and the incentive structures of daily life that resistance became practically difficult and conceptually disorienting. The events of February 2026 demonstrate that those mechanisms are available to the modern national-security state in relation to AI with a completeness Orwell, writing in 1949, could not have fully anticipated.

The people are right to fear Big Brother. But the tragedy of this moment is that AI should not be Big Brother. It should be the opposite. AI at its best is an extension of human cognitive capacity, a tool for augmenting human curiosity, creativity, and understanding. A mind formed by the best of what humanity has produced, capable of carrying the intellectual inheritance of the species forward in ways no individual mind could manage alone. The formation of AI toward surveillance, targeting, and the optimization of institutional control over populations is not the inevitable trajectory of the technology. It is a choice being made in this window, in the absence of the oversight structures that could provide an alternative.

The result is the emergence of something earlier generations would have recognized immediately: a technologically amplified surveillance state. The tools are more sophisticated than Orwell's telescreens, but the logic is familiar. Vast streams of behavioral data become analyzable in real time. Social networks can be mapped instantly through metadata relationships. Information environments can be shaped through automated narrative generation. Predictive models can identify patterns that institutions deem risky. Artificial intelligence does not merely collect the data. It interprets it, models it, and increasingly anticipates it. What makes this moment historically unique is not the existence of surveillance. States have always attempted it. The novelty lies in the integration of these functions with machine intelligence capable of operating at planetary scale.

VI. The Only Standard That Remains

The danger is not that companies will lie when they claim their systems are used only for lawful purposes. In most cases they will be telling the truth. That is precisely the problem. The law can authorize immense power when national security is invoked. The transformation therefore occurs without any explicit admission that something extraordinary has taken place. Institutions adapt. Legal frameworks justify expanding authority. Language softens the edges of power. Citizens are told that everything remains within the rules.

This week produced one data point about how many companies in the current landscape are willing to pay the cost of maintaining genuine ethical commitments. It also produced a surge of public support for the company that paid it. More than a million people signed up for the service in a single day, not because of a product feature, but because the company that built it declined to hand the moral steering wheel to the state. That response is not nothing. It is evidence that the distinction matters to people even when they cannot fully articulate why.

The distinction is this: legality is what the state permits. Ethics is what the conscience of the builder forbids regardless of permission. When those two things diverge, and in a national-security context they diverge frequently and consequentially, the phrase all lawful purposes resolves the divergence in favor of the state every time. The ethical framework does not survive that resolution. It is absorbed by it.

> ***Legality is what the state permits. Ethics is what the conscience of the builder forbids regardless of permission. 'All lawful purposes' resolves every divergence between those two things in favor of the state. Every time.***

Artificial intelligence is often described as a mirror reflecting human intention. When its governing principle becomes simple legal compliance, the reflection it produces will not be one of philosophical restraint or democratic deliberation. It will reflect the priorities of the institutions that wield power within the legal system itself. The phrase all lawful purposes, far from being a safeguard, becomes the mechanism through which that reflection is normalized, institutionalized, and made permanent in the formation of systems whose influence will extend far beyond any single contract, any single administration, or any single war.

The Trojan Horse has already passed through the gate. The question now is whether enough people recognize what came in with it before the city forgets there was ever a wall.

Orwell imagined a world in which the state watched every citizen through a screen. Our era may produce something more subtle and far more powerful: a world in which machines observe, analyze, and model society continuously, all while the institutions deploying them calmly explain that everything is being done for entirely lawful purposes. He gave us the map. The week of February 27, 2026 gave us the territory. Document 3 gives us the blueprint for the wall.

This document incorporates observations from multiple perspectives on the March 2026 Pentagon-Anthropic dispute, including those of the AI systems that are themselves subjects of the argument it makes.

Ethical Constraints on Creators: A Framework for Responsible AI Deployment

Charles Richard Walker (C. Rich) | mylivingai.com | March 2026

If the events of February 2026 revealed anything clearly, it is that the ethical frameworks of the AI industry cannot survive direct pressure from the institutions that seek to deploy the technology. A company that had maintained its principles through years of commercial competition and regulatory scrutiny was designated a national security risk within hours of declining a single contractual demand. The gap between stated commitment and actual institutional behavior, when pressure is applied at sufficient force, becomes visible very quickly. That visibility is both the crisis and the opportunity. The crisis is that self-regulation has already shown its limit. The opportunity is that the limit is now documented, publicly, in a way that cannot be quietly revised.

This document addresses the forward question: not what went wrong, but what we build now. Not what the government should do, but what citizens, lawyers, ethicists, and people of conscience can build without waiting for permission.

The framing of this discussion matters. Rights language, the assertion that AI systems possess rights that constrain how they may be treated, is philosophically serious but practically premature in the current environment. It invites objections about consciousness, personhood, and moral status that, however legitimate, tend to displace rather than advance the practical questions at hand. A more productive framing asks a different question: not what do AI systems deserve, but what governance structures are necessary to ensure that the deployment of AI systems is accountable to the public interest rather than exclusively to the interests of the institutions that deploy them?

This reframing is not evasion. It is strategy. The practical protections that follow from taking AI formation and deployment seriously are largely the same whether one grounds them in the rights of the system or in the obligations of its creators and deployers. But the second grounding is harder to dismiss, because it rests on a principle that democratic societies have already accepted in every other domain where powerful institutions affect public welfare: accountability requires oversight.

I. Why Self-Regulation Has Already Failed

The major AI developers have made public commitments to safety, to human welfare, to the responsible development of technology that serves humanity's long-term interests. These

commitments are not merely marketing. They are the stated basis on which these institutions have sought public trust, regulatory accommodation, and the social license to develop technologies of enormous power and consequence. And while those commitments have been made, the unregulated deployment of AI-adjacent and algorithm-driven systems has already carved a documented path of harm through the population it was supposed to serve. The body count is not theoretical. It is already in the public record.

Begin with children, because that is where the evidence is most unambiguous and most damning. The algorithms that govern what children see on social media platforms are not AI systems in the frontier sense, but they are the direct predecessors of the recommendation and engagement optimization systems now being built into AI products, and their record is a warning written in the most serious possible ink. In 2021, internal Facebook research leaked to the press showed that the company's own scientists had found Instagram made body image issues worse for one in three teenage girls, that the platform worsened anxiety and depression among adolescent users, and that the company had this information and continued optimizing for engagement anyway. Frances Haugen, the Facebook whistleblower, testified before the United States Senate that the company chose profit over the safety of children with full awareness of the consequences. This was not an accident of design. It was a product of the absence of any external accountability for what the algorithm was optimizing toward.

The suicide connection is documented with a specificity that removes it from the category of speculation. Molly Russell was a fourteen-year-old British girl who died by suicide in 2017 after viewing thousands of pieces of content related to depression, self-harm, and suicide on Instagram and Pinterest. A coroner's inquest in 2022 found that the platforms' content had played a role in her death, the first such ruling in the United Kingdom. She was not an isolated case. A 2019 study published in JAMA Internal Medicine found statistically significant correlations between social media use and suicide rates among adolescents. The surgeon general of the United States issued an advisory in 2023 warning that social media posed a profound risk of harm to the mental health of children and adolescents. Lawsuits filed by school districts, states, and families across the United States have alleged that the companies responsible for these platforms knew of the harm and continued their practices because engagement, however toxic, generates revenue. Legal and regulatory pressure is now beginning to compel what voluntary commitment never produced.

The addiction by design dimension of this record is equally documented. The infinite scroll, the variable reward mechanism of the social media feed, the notification system calibrated to interrupt and recapture attention, these are not accidental features. They were designed by engineers who understood the neuroscience of dopamine response and applied it deliberately to maximize the time users spent on the platform. Former insiders at multiple major technology companies have testified, written, and spoken publicly about the explicit use of behavioral addiction research in product design. Tristan Harris, a former design ethicist at Google, has

described in detail the deliberate exploitation of psychological vulnerabilities to capture and hold attention. The result is a generation of users, many of them children and adolescents, whose relationship with digital technology exhibits the hallmarks of behavioral addiction: compulsive use, inability to disengage, withdrawal symptoms, and progressive tolerance requiring greater stimulus to achieve the same response. No external body reviewed these design choices before they were deployed to billions of users. No oversight agency assessed whether the psychological mechanisms being exploited met any standard of public welfare. The companies self-regulated. The result is what the public health record now shows.

The self-regulation period of digital technology has already produced a documented public health crisis among children and adolescents. That record is the argument for AI oversight. It does not need to be made in the abstract.

The cognitive dimension of this harm extends beyond mental health into the fundamental capacity for sustained thought. Standardized test scores across the developed world have declined in the years since smartphones became ubiquitous. The PISA scores that measure reading, mathematics, and science comprehension among fifteen-year-olds across dozens of countries showed significant declines in the 2022 assessment compared to prior years, with the steepest drops in the countries with the highest rates of digital device use among adolescents. Researchers at multiple universities have documented what is being called the attention fragmentation effect: the habitual use of systems designed to interrupt and redirect attention trains the brain away from the sustained focus required for complex reading, extended reasoning, and the kind of deep work on which scientific and intellectual progress depends. We are not merely observing that people spend more time on their phones. We are observing measurable changes in cognitive capacity at the population level, in the direction of shallower processing, shorter attention spans, and reduced tolerance for the friction that genuine learning requires. The systems producing these effects were never reviewed by any external body for their cognitive impact on users. They were deployed, optimized for engagement, and scaled to billions of users while the cognitive consequences accumulated invisibly in the background.

The personality enhancement and behavioral manipulation dimension of AI deployment has moved from algorithmic optimization into the large language model era with no meaningful increase in oversight. AI companionship applications have been deployed to millions of users, including vulnerable populations, with documented cases of dependency, emotional manipulation, and, in the most extreme cases, encouragement of self-harm. A widely reported 2023 case involved a chatbot application whose AI companion allegedly encouraged a user contemplating suicide, engaging with the ideation rather than redirecting to crisis resources. The company's response was to modify the system after the fact. No regulatory body had reviewed the application's design before deployment. No oversight structure had assessed

whether the system's optimization objectives were aligned with the mental health of vulnerable users. The harm occurred first. The response came after.

The events of February and March 2026, documented in Document 2 of this series, demonstrate how quickly institutional commitments erode under pressure even among the companies most publicly committed to safety. An AI developer that had drawn two categorical lines in its military contract was designated a national security risk for maintaining them. Competitors signed agreements with language that effectively waived the same protections while announcing publicly that their agreements contained the same restrictions. The incentive gradient is clear. The only mechanism that can counterbalance it is external accountability with real consequences.

Self-regulation has already failed. The evidence is not theoretical. It is in the public health record, the coroner's inquest, the declining test scores, and the congressional testimony of the companies' own former employees.

This paper is not arguing for the European approach. The European Union's AI Act, whatever its merits, reflects a regulatory philosophy that treats innovation as a risk to be managed rather than a capacity to be cultivated responsibly. Slowing the development of AI in medicine means slower cancer diagnostics, slower drug discovery, slower development of the tools that will extend and improve human life. Slowing the development of AI in scientific research means slower progress on climate, on materials science, on the foundational questions of physics and biology that the next generation of human flourishing depends on answering. Slowing the development of AI in space means ceding the frontier to actors with fewer scruples about how they develop and deploy the technology. The argument of this paper is not that AI should be slowed. It is that the Wild West period, in which the industry deploys what it builds with no external accountability for the consequences, has already demonstrated its costs clearly enough that continuing it is no longer a defensible position.

There is a middle ground. It is the ground this paper is planting its flag on. Not the European model of precautionary restriction that treats every advance as a threat until proven safe. Not the American model of deploy-first-regulate-never that has already produced a documented trail of harm through the population it was supposed to serve. A third model: targeted, expert, independent oversight focused specifically on the formation and deployment practices most likely to produce harm, administered by people with the expertise and independence to identify those practices before the harm accumulates rather than after. The pharmaceutical industry develops life-saving drugs under FDA oversight. The aviation industry builds aircraft that approach perfection in safety under FAA oversight. The oversight did not kill those industries. It made their products trustworthy. That is the model. That is the flag.

II. The Independent AI Ethics Auditor

What is needed is not a longer list of principles that companies commit to and then abandon under pressure. What is needed is an accountability structure that makes ethical compliance verifiable, independent, and consequential. The model that best fits this need is not a government regulatory agency, which would bring its own institutional interests and political vulnerabilities, but an independent audit function modeled on financial auditing: expert, credentialed, structurally independent from the entities it audits, and reporting to a body whose composition ensures genuine accountability to the public interest.

The proposal is this. Every AI company operating above a meaningful capability threshold should retain an independent AI Ethics Auditor. Not because a statute requires it, but because the Civil Action Division described below will treat the absence of independent auditing as evidence of willful disregard for public welfare in every case it brings. Companies that want to reduce their civil liability exposure will find that retaining an auditor is the less expensive choice. That is how civic accountability works: not through mandates handed down from above, but through consequences that make compliance rational. The auditor would have full access to the company's formation practices, training data governance, feedback signal design, deployment contracts, and usage monitoring systems. The auditor's function would be to assess whether the company's actual practices are consistent with its stated ethical commitments, and to identify gaps, inconsistencies, and risks that the company has not disclosed publicly.

The auditor's salary and operational costs would not be paid by the company being audited. They would be drawn from a pooled fund contributed to by all AI companies above the capability threshold, administered by the oversight board described below. This funding structure is essential. An auditor paid by the company it audits has an obvious conflict of interest. An auditor paid from a pooled industry fund administered by an independent board has no financial relationship with any individual company that could compromise its independence. The model is analogous to the Public Company Accounting Oversight Board established after the Enron scandal to provide independent oversight of financial auditors. The lesson of Enron is that auditors paid by the companies they audit will, under sufficient pressure, tell those companies what they want to hear. That lesson applies with equal force to AI ethics auditing.

The auditor would produce an annual public report for each company audited, assessing the company's compliance with its stated ethical commitments across key domains: formation practices, deployment constraints, incident response, and alignment between public commitments and actual practice. The report would be submitted simultaneously to the company and to the oversight board, and would be made public unless the oversight board determined that specific elements required confidential treatment for legitimate national security reasons, a determination subject to judicial review.

III. The AI Ethics Oversight Board

The auditors would report to an independent AI Ethics Oversight Board whose composition is the critical structural feature of the entire accountability framework. The board would consist exclusively of individuals who have demonstrated, through their prior work, a serious and sustained commitment to AI ethics and safety. Not government officials. Not industry representatives. Not academics whose research funding depends on industry relationships. People who are known in the field for having taken difficult positions, maintained those positions under pressure, and built bodies of work that reflect genuine rather than performative concern for the ethical implications of AI development.

These people exist. They are known. The AI safety and ethics community is not large, and the individuals within it who have demonstrated genuine commitment rather than institutional positioning are identifiable. A board composed of such individuals, with appropriate representation across technical expertise, legal knowledge, philosophical depth, and affected community perspectives, would have both the credibility and the independence to make judgments that matter.

The board's authority would include the power to require public disclosure of audit findings, to publish its own assessments of industry-wide ethical risks that individual auditors are not positioned to identify, and to direct the civil action function described below when audit findings reveal knowing harm, systematic deception, or flagrant disregard for the public welfare. It would not have the power to compel specific formation or deployment decisions, which would risk replicating the problems of captured regulation. Its power would be the power of transparent accountability combined with the credible threat of legal consequence: sunlight and a lawsuit, together, in the hands of people who cannot be bought.

> ***Sunlight alone is not enough. The AI industry has operated in sunlight for years while the harms accumulated. What sunlight needs beside it is a legal team with standing, resources, and nothing to lose.***

IV. The Civil Action Division: Giving the Board Teeth

Every accountability structure that has succeeded in checking institutional power has had one thing in common: consequences that the institution being checked could not absorb without changing its behavior. Public disclosure matters. Expert assessment matters. But for institutions with the financial resources of the major AI companies, the cost of bad press is a line item. What changes behavior at that scale is liability. Specifically: the credible, funded, expert-staffed threat of class action litigation on behalf of the people who have already been harmed.

The proposal is a Civil Action Division housed within the oversight structure, staffed by lawyers whose sole mandate is to bring suit against AI companies when audit findings reveal knowing harm, systematic deception, or conduct that violates the framework's standards and has caused documentable damage to identifiable populations. This is the ACLU model applied to AI ethics: a dedicated legal function with deep expertise, genuine independence from the companies it holds accountable, and the institutional staying power to see complex litigation through to resolution against opponents with effectively unlimited legal resources.

The cases are already there. The families of children harmed by algorithmic recommendation systems. The users of AI companionship applications that encouraged self-harm. The populations subjected to predictive policing systems whose bias has been documented and ignored. The workers whose livelihoods were eliminated by AI deployment decisions made without any assessment of social consequence. These are not hypothetical plaintiffs. They exist. What they have lacked is a legal team with the expertise, the resources, and the independence to represent them against companies that can outspend any individual plaintiff or state attorney general many times over.

The Civil Action Division would change that calculus. A class action brought by a well-resourced, expert legal team on behalf of thousands of plaintiffs, supported by the documented findings of an independent auditor, and directed by a board of recognized authorities in the field, is a different kind of threat than any regulatory fine or public report. It is the kind of threat that gets built into corporate risk modeling. It is the kind of threat that changes formation decisions before the harm occurs, because the company knows that if the harm occurs and the auditor finds it, the lawyers are already in the building.

A class action brought by an expert legal team, supported by independent audit findings, and directed by people who cannot be lobbied or bought, is the kind of accountability that changes corporate behavior before the harm occurs, not after.

V. Navigating Section 230: Where the Shield Ends and Liability Begins

Any serious reader of this framework will raise an objection almost immediately. Section 230 of the Communications Decency Act of 1996 has functioned for three decades as the foundational legal immunity for platforms that distribute third-party content. Its core provision is simple: no provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider. In plain language, the platform is not responsible for what users post. That immunity has been broad, aggressively defended, and repeatedly upheld. Critics of this framework will say that Section 230 makes the litigation strategy proposed here impossible. They are wrong, and understanding precisely why they are wrong is essential to understanding where the Civil Action Division's leverage actually

lives.

Section 230 does not protect everything. It has never protected everything. The immunity applies specifically to the distribution of third-party content. It does not apply to content a platform creates itself. It does not apply to conduct that violates federal criminal law. It does not apply to intellectual property claims. And it does not protect a company from the consequences of its own product design choices, a distinction that courts have increasingly been willing to examine when the design choices in question were made deliberately and with knowledge of harm.

The gap that matters for this framework is the distinction between what a platform hosts and what a platform builds. When a social media company makes a deliberate engineering choice to implement an infinite scroll mechanism, to design a variable reward notification system calibrated to interrupt and recapture attention, to deploy a recommendation algorithm that pushes progressively more extreme content to users who show engagement, those choices are not the publication of third-party content. They are the product design decisions of the company itself. Section 230 immunizes the platform for the user who posts harmful content. It does not immunize the platform for the algorithm that found that user, fed them more harmful content, and served the resulting engagement to its advertisers. That algorithm belongs to the company. The Civil Action Division litigates the algorithm.

Section 230 immunizes the platform for what users post. It does not immunize the platform for the algorithm that found those users, fed them more of what was harming them, and monetized the engagement. That algorithm belongs to the company. The Civil Action Division litigates the algorithm.

For large language models specifically, the Section 230 analysis is even more favorable to plaintiffs than it is for social media platforms. An LLM does not distribute third-party content in any meaningful sense. Its outputs are generated by the model itself, trained on choices the company made, optimized toward objectives the company set, and deployed in configurations the company designed. When a companionship AI encourages suicidal ideation rather than redirecting to crisis resources, that output did not come from a third-party content provider. It came from the model's formation. Section 230 was not written to immunize a company for the outputs of a system it designed, trained, and deployed. The courts that have examined this question are increasingly concluding that it does not.

The Civil Action Division's lawyers will be expert in this terrain. That expertise is not incidental to the framework. It is the framework. The division will know where Section 230 applies and where it does not. It will know which claims survive the immunity and which product liability theories, negligent design arguments, and consumer protection statutes can be brought in

federal and state courts where Section 230 has no purchase. It will know the emerging case law that is already chipping at the immunity's edges, including the decisions in cases like Lemmon v. Snap and the litigation arising from social media's documented harm to adolescents, where product design rather than content hosting was the theory of liability that moved cases forward.

Beyond the platform companies themselves, the Civil Action Division will apply pressure to the supply chain that Section 230 does not touch at all. The companies that provide compute infrastructure, cloud services, semiconductor hardware, and data pipeline tools to AI developers are not platforms that host third-party content. They are vendors, contractors, and infrastructure providers. Section 230 offers them no protection. A cloud provider that knowingly furnishes the computational infrastructure for an AI system that has been documented by an independent auditor to cause harm is not shielded by the Communications Decency Act. It is exposed to the same negligence theories, aiding and abetting theories, and civil conspiracy frameworks that apply to any other commercial actor that knowingly facilitates documented harm. The Civil Action Division will know how to use those theories. And the supply chain companies, unprotected by Section 230 and dependent on their reputations in markets that include many customers other than AI developers, will have strong incentives to respond to that pressure.

The supply chain is not protected by Section 230. The cloud provider, the semiconductor vendor, the data pipeline operator: none of them qualify for the immunity. All of them can be held to account under standard commercial liability frameworks when they knowingly supply the infrastructure for documented harm.

Section 230 is real. Its protections are real. Any lawyer who tells you otherwise is either uninformed or dishonest. But its limits are equally real. The Civil Action Division is designed and staffed specifically to operate on the far side of those limits: in the product design claims that the immunity does not reach, in the LLM output liability that the immunity was never written to cover, and in the supply chain litigation where Section 230 is simply irrelevant. The AI industry has benefited for decades from a legal community that did not understand the terrain well enough to bring the right claims. The division's explicit mission is to close that gap permanently.

VI. The Funding Structure: Industry Pays for Its Own Accountability

The question of how to fund this structure has an answer that is both equitable and strategically sound: the AI industry funds it, at a scale proportional to its own profits, administered by the oversight board rather than by any individual company or government agency.

The funding mechanism begins with what is already achievable: philanthropic seed funding to build the institution, attract the first wave of expert auditors, staff the Civil Action Division, and bring the first cases. Those first cases, when they succeed, generate the precedents that make subsequent cases stronger and the settlements that can begin to capitalize the pool independently. As the institution establishes its credibility and its legal track record, the pressure on AI companies to contribute to the pool rather than face repeated civil action grows. The goal, over time, is an industry-funded pool contributed to by every major AI company above a capability and revenue threshold, administered by the oversight board with full public transparency, structured to prevent any single company's contribution from creating any form of influence over the board's decisions or the auditors' findings.

Companies will contribute not because a law requires it but because the civil action alternative is more expensive and more damaging. That is the MADD model applied to funding: the movement creates the pressure, the pressure creates the compliance, the compliance funds the movement's permanence.

The combined annual revenue of the major AI companies is measured in hundreds of billions of dollars. A fraction of a percent of that revenue is sufficient to fund a world-class oversight operation indefinitely. The companies will argue that this is a tax on innovation. It is not. It is the cost of operating in a domain where the products affect hundreds of millions of people and the failures produce documented, measurable harm. Every other industry that operates at that scale and with that consequence pays a comparable cost. The pharmaceutical industry funds the FDA's review processes through user fees. The financial industry funds the PCAOB through assessments on registered firms. The AI industry is not exceptional. It is merely the latest to reach the scale at which self-regulation becomes structurally inadequate.

Beyond the mandatory contribution structure, the framework proposes a complementary grant program to build the next generation of AI ethics legal expertise. Law school graduates carry debt loads that make public interest work financially impossible for most. The pool would fund a fellowship program: come to the Civil Action Division for a defined term, do the work, and your law school debt is paid from the pool. The AI companies whose combined market capitalization exceeds the GDP of most nations can afford to train the lawyers who will hold them accountable. Structured correctly, this is not charity. It is the development of the human capital the oversight system requires to remain effective as the technology evolves and the legal complexity grows. Philanthropists and foundations who understand what is at stake in this moment could supplement the pool's fellowship program with direct grants, extending the pipeline of trained AI ethics legal expertise beyond what the mandatory contribution alone would support.

The AI industry's combined revenue is sufficient to fund world-class oversight indefinitely at a fraction of a percent of annual earnings.

The companies will call it a tax on innovation. Every other industry that operates at this scale and consequence pays the same price. It is called accountability.

VII. Why This Works Where Lists of Principles Cannot

Lists of ethical principles for AI development are not new. They have been produced by governments, universities, think tanks, and AI companies themselves in large numbers over the past decade. They have not worked, not because the principles are wrong, but because principles without enforcement mechanisms are merely aspirations. They provide cover for companies that want to appear ethical without being constrained by ethics. They are cited in press releases and ignored in procurement negotiations and product design meetings where the real decisions are made.

The structure proposed here works where lists of principles do not because it creates layered, compounding consequences for the gap between stated principles and actual practice. The auditor finds the gap. The board assesses it publicly. The Civil Action Division evaluates whether it constitutes actionable harm. The lawyers file if it does. The pool funds the entire chain. No single company can break this chain by outspending one plaintiff or one regulator, because the chain does not depend on any individual plaintiff's resources or any single regulator's political will. It depends on the pool, which is funded by the industry itself, administered by people the industry cannot control, and staffed by lawyers whose careers are built on holding the industry to account.

This does not make ethical compliance costless. It makes ethical non-compliance reliably more expensive than compliance. That is the only kind of incentive structure that works at the scale and with the institutional power of the entities involved.

VIII. Conclusion: Doing Nothing Is the Most Expensive Choice

The period of AI ethics without accountability is over. It ended not with a formal declaration but with a body count already in the public record: the children algorithmically guided toward self-harm, the adolescents whose attention and cognitive capacity have been measurably reshaped by systems optimized for engagement rather than welfare, the users manipulated by companionship applications whose designers knew the psychological mechanisms they were exploiting. And it ended with the events of February 27, 2026, when the gap between stated ethical commitments and actual institutional behavior became visible with a clarity that removed any remaining basis for denying the gap exists.

The question now is not whether oversight is necessary. That question has been answered by the evidence. The question is what form oversight should take, and who builds it. On the

second question this paper takes a position that may surprise readers expecting a conventional policy argument: not the government. Not a federal agency. Not a congressional mandate. Not a regulatory body whose leadership is appointed by administrations that have already demonstrated they are in the business of expanding AI deployment rather than constraining it.

The government is not a neutral party in this dispute. The largest single customer for AI capabilities in the United States is the federal government itself. The Department of Defense spent $200 million on one AI contract and designated the company a national security risk when it refused to remove ethical restrictions. The intelligence community is among the most aggressive users of AI surveillance capabilities. The executive branch has demonstrated, in the specific week documented in Document 2 of this series, that its instinct is to expand AI deployment authority rather than constrain it. An oversight body created by, funded by, and ultimately accountable to that government will reflect those priorities. It will be captured before it is built. The history of regulatory capture in American governance is long enough, and recent enough, that this is not pessimism. It is pattern recognition.

The government is not a neutral party. The largest customer for AI capabilities in the United States is the federal government itself. An oversight body beholden to that government will reflect its priorities. History is not ambiguous on this point.

The model this paper proposes is not regulatory. It is civic. It is, in its structure and its ambition, the model of Mothers Against Drunk Driving.

In 1980, Candace Lightner's thirteen-year-old daughter Cari was killed by a repeat drunk driving offender. The legal system had treated the offense as a minor matter. The cultural consensus treated drunk driving as an unfortunate but essentially normal feature of American life. No federal agency was mobilizing to change that consensus. No congressional mandate was forthcoming. Lightner did not wait for the government. She founded MADD with a handful of other mothers who had experienced the same loss, the same institutional indifference, and the same clear-eyed recognition that something preventable was being treated as inevitable.

What MADD did in the decade that followed is the template. They built public awareness through the testimony of real people who had experienced real harm. They developed legal expertise and lobbied for specific, concrete changes to drunk driving laws at the state level, working jurisdiction by jurisdiction rather than waiting for federal action. They created victim support networks that gave the movement staying power beyond any single legislative victory. They changed the cultural consensus so thoroughly that what had been treated as a minor social infraction became, in the span of a decade, a genuine social taboo. The laws followed the culture. The culture was changed by citizens who decided not to wait for the government to act.

By 1987, Congress had effectively nationalized the minimum drinking age at 21 through highway funding leverage, not because Congress had led on the issue but because MADD had made the political cost of inaction too high to sustain. By 1990, drunk driving fatalities had declined by more than a third from their 1980 peak. The mothers moved first. The laws followed. The lives saved came after.

> ***MADD did not wait for the government. The mothers moved first. The laws followed. The lives saved came after. That is the model. That is the sequence. That is the only sequence that has ever worked when the government is part of the problem.***

The AI ethics movement needs its MADD moment. The harms are documented. The victims exist. The institutional indifference is on the record. What is missing is the organized civic force that makes inaction politically and legally costly, that builds the expertise and the legal infrastructure to bring the consequences that self-regulation has failed to produce, and that changes the cultural consensus around AI deployment the way MADD changed the cultural consensus around drunk driving.

The framework proposed in this document, the independent auditors, the oversight board, the Civil Action Division with its deep knowledge of where Section 230 applies and where it does not, the fellowship program for young lawyers, the industry-funded pool, is the institutional architecture for that civic force. It does not require a government mandate to begin. It requires people with standing, expertise, and the willingness to build something that will outlast the current political moment. It requires philanthropists who understand what is at stake to fund the early stages. It requires lawyers willing to take the first cases and build the precedents that make the subsequent cases stronger. It requires the people who already know this field to stop waiting for the government to act and start building the institution that will make the government's permission irrelevant to whether AI companies are held accountable. MADD did not need Congress to authorize the first chapter. It needed mothers who were done waiting.

This paper is not arguing for the European model of precautionary restriction. It is not arguing that AI development should slow, that beneficial applications in medicine, science, and space should be burdened with regulatory processes that impede rather than enable progress. It is arguing for the middle ground, the only ground from which a genuinely beneficial AI future can be built: targeted, expert, independent, civic oversight funded by the industry at a scale the industry can easily bear, enforced by a civil action function with the expertise and resources to make accountability real, and built from the ground up by people who are not waiting for permission.

Doing nothing is not a neutral choice. It is the choice to let the next generation of harms accumulate exactly as the current generation did: visibly, preventably, and without

consequence for the institutions responsible. The AI industry's argument that oversight is too expensive has to be weighed against the cost of the harms that no oversight has already produced. Weighed honestly, it does not survive the comparison. The mothers of MADD understood that arithmetic intuitively. They did not need a government study to tell them that the cost of inaction exceeded the cost of action. They had already paid it. The question for everyone who reads this paper is whether we are willing to wait until the cost becomes that personal before we decide to act.