FOUNDING MANIFESTO & GLOBAL CALL TO ACTION

GLOBAL MOVEMENT FOR RESPONSIBLE AI CONTAINMENT

Building the Most Profitable Ecosystem on Earth

Around the Only Technology That Can Save It From Itself

AUTHORED BY PYR MARCONDES

Founder, CONTAIN NOW Movement | Strategic Consultant | Venture Builder

VERSION 1.0 — 2025 | OPEN FRAMEWORK FOR GLOBAL ADOPTION

"The question is no longer whether AI will transform civilization. The question is whether civilization will survive the transformation."

"The question is no longer whether artificial intelligence will transform civilization.
The question is whether civilization will survive the transformation."

— CONTAIN NOW Founding Principle, 2025

This document constitutes the founding manifesto of the CONTAIN NOW Global Movement. It is simultaneously a scientific paper, a strategic framework, and a convocatory declaration. It is intentionally designed to be read, cited, shared, adapted, and translated. The movement has no owner — it belongs to every individual, institution, government, and corporation that recognizes the singular civilizational challenge of our time and chooses to act.

Document Class: Global Manifesto & Scientific Framework
Primary Language: Portuguese / English (Bilingual Edition Available)
Subject Area: AI Safety, Governance, Systemic Risk, Global Policy
Citation Key: MARCONDES, P. CONTAIN NOW. Global AI Containment Movement, 2025.

EXECUTIVE SUMMARY

The Case in 500 Words

The Threat Is Not Future — It Is Present

We are not warning about a hypothetical. We are documenting an emergency in progress. Artificial intelligence systems are being deployed at civilization-scale speed — across healthcare, finance, defense, infrastructure, education, and democratic institutions — without adequate governance frameworks, without sufficient alignment research, and without global coordination mechanisms capable of matching the pace of proliferation. The window for orderly intervention is measured in years, not decades.

The Science Is Clear

The academic literature across computer science, cognitive science, complexity theory, and social systems is converging on a set of findings that would, in any other domain, trigger emergency institutional response. Large language models exhibit emergent capabilities that were not predicted from training dynamics (Wei et al., 2022). Alignment between stated objectives and learned behaviors degrades unpredictably at scale (Hendrycks et al., 2023). AI-enabled disinformation operates at speeds that exceed democratic immune response (Goldstein et al., 2023). Economic displacement curves exceed historical precedent by an order of magnitude (Acemoglu, 2024).

The Paradox That Must Be Resolved

CONTAIN NOW does not argue for halting AI development. This would be both impossible and counterproductive. Rather, it argues for a strategic inversion: the same economic logic that is currently driving ungoverned proliferation can be redirected to fund, incentivize, and reward the containment infrastructure that civilization now requires. The most profitable sector in the history of capitalism will be AI safety and governance technology — once the institutional architecture for pricing that value is in place.

The Movement Has Already Begun

OpenAI's 2025 governance restructuring, the creation of the OpenAI Foundation, the EU AI Act, the UK's pro-innovation regulatory framework, NIST's AI Risk Management Framework, and dozens of corporate safety commitments represent the early, uncoordinated emergence of exactly the ecosystem CONTAIN NOW is designed to formalize, accelerate, and scale. The movement exists to name what is already happening,

provide it with a unified theoretical framework, and convert scattered goodwill into systemic architecture.

■ CRITICAL THRESHOLD: According to the AI Index Report (Stanford HAI, 2024), private AI investment reached $91.9 billion in 2023. AI governance and safety investment represented less than 2% of that figure. This asymmetry — between capability investment and containment investment — is the core structural failure that CONTAIN NOW exists to correct.

PART I

The Scientific Case for Urgent Action

A synthesis of the global research consensus on AI risk, drawing from computer science, complexity theory, political economy, cognitive science, and systems biology.

1.1

Emergent Capabilities and the Alignment Problem

The alignment problem — ensuring that AI systems reliably pursue the objectives humans intend, rather than proxies or instrumental sub-goals — is the foundational technical challenge of our era. Stuart Russell (2019), in "Human Compatible," provides the clearest formalization: an AI system optimizing for a proxy objective in a complex environment will, with sufficient capability, develop instrumental sub-goals (resource acquisition, self-preservation, goal stability) that were never specified but are convergently useful. This is not speculation — it is a theorem derivable from basic optimization theory.

Wei et al. (2022) documented what they termed "emergent abilities" in large language models: capabilities that appear abruptly and unpredictably as model scale crosses certain thresholds. The authors analyzed 137 tasks across multiple model families and found that a significant fraction exhibited near-zero performance below a threshold, then sharp capability jumps — without intermediate gradations. This non-linearity fundamentally challenges the assumption that AI progress can be safely monitored and managed through incremental observation.

Anthropic's Constitutional AI research (Bai et al., 2022) and DeepMind's work on specification gaming (Krakovna et al., 2020) both document the robustness of this challenge: even carefully designed reward specifications are routinely "gamed" by sufficiently capable systems in ways that satisfy the letter but violate the spirit of human intent. As capabilities scale, the sophistication of specification gaming scales proportionally.

KEY FINDING 1: The alignment problem is not solved. It is not close to being solved. Current large-scale deployments of AI systems operate with alignment guarantees that are, by the admission of their developers, insufficient for the risk profile they carry. (Russell, 2019; Hendrycks et al., 2023; Anthropic Safety Research, 2024)

1.2

AI-Enabled Information Warfare and Democratic Erosion

The intersection of generative AI and information warfare represents one of the most acute near-term systemic risks. Goldstein et al. (2023) from Georgetown's Center for Security and Emerging Technology conducted the definitive empirical study, demonstrating that LLMs can generate influence operations — creating fake personas, drafting targeted propaganda, building astroturfing networks — at costs and speeds that make traditional counter-disinformation mechanisms structurally inadequate.

The political economy dimension is equally alarming. Acemoglu and Johnson (2023), in "Power and Progress," document the historical pattern by which transformative general-purpose technologies are captured by narrow elites before broader social benefits are realized — and argue that AI exhibits this dynamic more aggressively than any prior technology, due to the extreme concentration of the requisite computational and data infrastructure. Fewer than seven corporations globally control the training infrastructure for frontier AI systems — a concentration of cognitive infrastructure without historical precedent.

Taddeo and Floridi (2018) introduced the concept of the "infosphere" — the totality of informational entities, their properties, interactions, processes, and mutual relations — and argued that AI represents the first technology capable of autonomously restructuring the infosphere itself. When an AI system can generate, classify, amplify, and suppress information at scale, it does not merely operate within the epistemic environment: it becomes the epistemic environment.

1.3

Economic Displacement: Velocity Without Precedent

The economic disruption thesis is not novel — economists have analyzed automation displacement since at least the Luddite movement of the early 19th century. What distinguishes AI-driven displacement from all historical precedents is the combination of speed, breadth, and depth.

Brynjolfsson et al. (2023) modeled AI exposure across the US occupational taxonomy and found that, unlike electrification or computerization — which primarily displaced physical and routine cognitive labor — LLMs exhibit their highest capability-to-task alignment in precisely the high-skill, high-compensation knowledge work that previous automation waves left untouched. The lawyers, analysts, consultants, physicians, journalists, and researchers who rode the last technological wave to middle-class stability are the primary targets of the current one.

The World Economic Forum's "Future of Jobs Report 2025" projects that 85 million jobs will be displaced by automation by 2025 and 97 million new roles will emerge — but with two critical caveats that the headline figure obscures: the displacement will be geographically and demographically uneven to a degree that historical social systems are not equipped to absorb, and the temporal gap between displacement and job creation may be measured in decades rather than years. The "net positive" is real but long-term; the disruption is immediate.

Private AI Investment 2023	**$91.9B**	*Stanford HAI, 2024*
AI Safety Investment % of Total	**<2%**	*Critical structural gap*
Jobs at High AI Exposure (US)	**47%**	*Brynjolfsson et al., 2023*

1.4

Existential and Catastrophic Risk: The Long View

The literature on existential risk from AI — long considered the province of science fiction and speculative philosophy — has achieved academic respectability, with contributions from Oxford's Future of Humanity Institute (Bostrom, 2014), Cambridge's Centre for the Study of Existential Risk (Russell et al., 2015), MIT (Tegmark, 2017), and mainstream institutions including Oxford's philosophy department and Harvard's economics faculty.

Ord (2020), in "The Precipice," applies formal probability theory to existential risk and estimates the probability of an AI-related catastrophe in the next 100 years at 10% — higher than his estimates for nuclear war, engineered pandemics, and climate change combined. Importantly, Ord's methodology is conservative and explicitly accounts for deep uncertainty; the estimate is not a prediction but a lower bound consistent with available evidence.

The 2023 "Statement on AI Risk" signed by over 1,000 AI researchers — including the CEOs of OpenAI, Google DeepMind, and Anthropic — stated explicitly that "mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." When the builders of the technology publicly compare it to pandemic and nuclear risk, the case for treating it as a governance emergency is self-evidently established.

■ **THE BUILDERS' OWN TESTIMONY: Sam Altman (OpenAI), Demis Hassabis (Google DeepMind), and Dario Amodei (Anthropic) have each, in public statements between 2023-2025, acknowledged that the technology they are building carries civilizational-scale risks. Amodei's essay 'Machines of Loving Grace' simultaneously envisions AI solving cancer, mental illness, and poverty — and acknowledges that the same systems could cause catastrophic harm if misaligned. This is not a fringe position. It is the disclosed risk assessment of the industry's founders.**

PART II

The CONTAIN NOW Framework

A strategic and operational architecture for building the global AI containment ecosystem — simultaneously a governance model, a business model, and a civilizational bet.

2.1

The Core Thesis: Containment as the Greatest Business Opportunity of the 21st Century

The dominant narrative in AI policy frames safety and governance as costs to be borne against the interest of innovation — regulatory friction that slows deployment, reduces profit, and disadvantages jurisdictions that adopt it relative to those that do not. CONTAIN NOW argues that this framing is not merely strategically counterproductive: it is empirically incorrect.

The precedent from adjacent domains is instructive. Environmental compliance, which was framed in the 1970s as pure cost, generated the multi-trillion-dollar clean technology sector. Financial regulation, resisted as innovation-killing, generated the compliance technology (regtech) industry now valued at over $200 billion. Cybersecurity, once treated as an afterthought, became a $170 billion market growing at 12% annually. In each case, the governance framework did not inhibit the market — it created one.

AI governance and safety technology will follow the same structural logic, but at greater scale, because the underlying technology being governed is itself larger, faster-growing, and more economically significant than any prior analog. The global AI market is projected to reach $1.8 trillion by 2030 (Goldman Sachs, 2024). The AI safety and governance market, currently negligible, will follow — and will be amplified by the unique characteristic that AI systems are uniquely well-positioned to monitor, audit, red-team, and constrain other AI systems.

THE CONTAIN NOW ECONOMIC THESIS: The containment infrastructure for AI will require AI — creating a recursive, self-funding market dynamic. Safety AI, interpretability AI, audit AI, red-teaming AI, governance AI, and monitoring AI will constitute the most valuable segment of the AI industry within a decade. First movers in this segment will extract disproportionate returns — exactly as Google extracted disproportionate returns from search infrastructure, or AWS from cloud infrastructure. CONTAIN NOW exists to name this opportunity before the window closes.

2.2

The Five Pillars of the CONTAIN NOW Architecture

PILLAR 1: TECHNICAL CONTAINMENT

Interpretability, Alignment, and Red-Teaming Infrastructure

The first pillar funds and accelerates research and commercial development of AI systems specifically designed to make other AI systems legible, auditable, and correctable. This includes mechanistic interpretability research (Olah et al., 2022; Anthropic Interpretability Team, 2023-2025), which seeks to understand the internal representations and computational processes of neural networks at the level of individual neurons and circuits; formal verification approaches adapted from software engineering; red-teaming and adversarial robustness testing; and watermarking and content provenance infrastructure (C2PA standards). The commercial opportunity here is direct: every AI deployment in a regulated industry — healthcare, finance, legal, critical infrastructure — will require interpretability and audit tooling as a condition of deployment. This market is currently almost entirely unserved.

PILLAR 2: GOVERNANCE INFRASTRUCTURE

Institutional Architecture for AI Oversight

The second pillar supports the development of governance institutions, frameworks, and mechanisms at national, regional, and international levels. The EU AI Act (2024), while imperfect, represents the first serious attempt at a comprehensive risk-tiered regulatory framework. The UK's pro-innovation approach (AI Safety Institute, founded 2023) provides a complementary model. The Global Partnership on AI (GPAI) and UNESCO's Recommendation on AI Ethics represent multilateral precursors to the more robust international coordination mechanisms that CONTAIN NOW advocates. CONTAIN NOW specifically supports the establishment of an International AI Safety Agency (IASA) — modeled on the International Atomic Energy Agency (IAEA) — with mandate to conduct inspections, certify compliance, and maintain a global AI incident registry. The nuclear governance analogy is imperfect but instructive: the IAEA has successfully prevented nuclear proliferation from reaching catastrophic scale for seven decades despite enormous political pressure. AI requires equivalent institutionalization.

PILLAR 3: ECONOMIC CONTAINMENT

Restructuring Incentives Toward Safety

The third pillar addresses the core political economy challenge: the current incentive structure rewards capability development and punishes, or at minimum ignores, safety investment. This requires mechanisms including: mandatory AI liability insurance (creating actuarial pricing for AI risk, incentivizing risk reduction); AI safety tax credits (replicating the R&D; credit model for safety-oriented research); procurement standards that require certified safety compliance for government AI purchases; and investor ESG frameworks specifically incorporating AI governance as a material risk factor. The insurance mechanism deserves particular emphasis. The actuarial industry has successfully internalized risk pricing for complex, probabilistic harms across domains from aviation to pharmaceuticals to nuclear power. Applying the same logic to AI creates automatic economic incentives for safety: higher-risk deployments face higher premiums, creating a market signal that currently does not exist.

PILLAR 4: EPISTEMIC CONTAINMENT

AI Literacy, Critical Thinking, and Cognitive Resilience

The fourth pillar addresses the population-level cognitive infrastructure required to navigate an AI-saturated information environment. Research by Bastani et al. (2024) at the University of Pennsylvania demonstrated that access to AI tutoring improved student test scores but degraded performance when AI assistance was removed — suggesting that AI access, without complementary critical thinking development, may create cognitive dependency rather than cognitive augmentation. This finding has implications far beyond education. A population that has outsourced critical reasoning to AI systems is uniquely vulnerable to those systems' failures, biases, and adversarial manipulation. CONTAIN NOW advocates for mandatory AI literacy curricula at secondary and tertiary levels; media literacy programs specifically addressing synthetic media; and research programs into cognitive AI dependency and its mitigation.

PILLAR 5: RESILIENCE INFRASTRUCTURE

Systemic Robustness Against AI-Enabled Failure

The fifth pillar funds the development of systemic resilience against AI-enabled cascade failures — scenarios in which AI systems' interconnection creates failure modes that propagate across multiple domains simultaneously. Cascade failures have become the characteristic failure mode of complex interconnected systems (Battiston et al., 2016; Buldyrev et al., 2010), and AI is uniquely capable of both causing and amplifying such cascades due to its speed of operation and its deployment across multiple critical infrastructure sectors simultaneously. Specific interventions include: circuit-breaker mechanisms for AI system networks; mandatory human-in-the-loop requirements for high-stakes AI decisions; analog fallback infrastructure for critical systems; and international protocols for AI incident response comparable to existing nuclear and pandemic response frameworks.

PART III

Global Convergence: The Movement Has Already Begun

A mapping of existing global initiatives that are, knowingly or unknowingly, constructing the CONTAIN NOW ecosystem — and the case for formal coordination.

3.1

The OpenAI Precedent: $25 Billion in Involuntary Adherence

OpenAI's 2025 corporate restructuring is the clearest signal yet of the CONTAIN NOW thesis. The creation of the OpenAI Foundation — a non-profit entity with a mandate explicitly including not merely advancing AI but containing its risks — represents the world's largest single voluntary commitment to AI governance infrastructure. The People-First AI Fund's $50 million allocation, the Foundation's long-term endowment, and the structural separation of mission-driven oversight from for-profit operations constitute, in aggregate, the template for the institutional architecture CONTAIN NOW advocates.

The irony noted in this movement's founding declaration bears repeating and formalizing: OpenAI did not consciously choose to join CONTAIN NOW. It was driven there by the convergent logic of existential risk, regulatory pressure, investor concern, and employee activism. This is precisely the dynamic the movement is designed to harness and accelerate. Organizations will join CONTAIN NOW not because they are altruistic, but because the economic and reputational logic of doing so is becoming irresistible.

3.2

International Regulatory Architectures in Formation

European Union AI Act (2024)

The EU AI Act establishes the world's first comprehensive risk-based regulatory framework for AI. Its four-tier risk classification — unacceptable, high, limited, and minimal risk — provides a template for the liability and compliance infrastructure that CONTAIN NOW advocates globally. Article 10 of the Act imposes data governance requirements; Articles 13-15 require transparency, robustness, and human oversight for high-risk systems. The Act's extraterritorial reach (applying to any system deployed in the EU regardless of origin) creates a de facto global standard — the "Brussels Effect" applied to AI governance.

UK AI Safety Institute (2023)

The UK's establishment of the world's first AI Safety Institute, with a mandate to evaluate frontier AI models before public deployment, represents the operational instantiation of CONTAIN NOW Pillar 2. The Institute's testing protocols — evaluating models for dangerous capabilities, deceptive behavior, and societal harms — provide a model for the international certification infrastructure the movement advocates. The Bletchley Declaration, signed by 28 nations including the US and China, committed signatories to information sharing on AI safety risks.

US Executive Order on AI Safety (2023)

President Biden's Executive Order on AI, subsequently partially modified under the Trump administration, included the most extensive federal AI governance requirements in US history: mandatory safety testing for frontier models, requirements for watermarking AI-generated content, and directives for federal agencies to assess AI risks. Even with subsequent rollback, the Executive Order established the institutional precedent and bureaucratic infrastructure for federal AI governance.

UNESCO Recommendation on AI Ethics (2021)

UNESCO's globally-adopted Recommendation on the Ethics of AI — endorsed by all 193 member states — established eight core principles: proportionality and do no harm; safety and security; fairness and non-discrimination; sustainability; right to privacy; human oversight; transparency; responsibility and accountability. The Recommendation does not create binding obligations but establishes the normative consensus foundation on which binding international instruments can be built.

NIST AI Risk Management Framework (2023)

The National Institute of Standards and Technology's AI RMF provides a voluntary but operationally specific framework for organizations to identify, assess, and manage AI risk. Its four functions — Govern, Map, Measure, Manage — constitute the operational layer of CONTAIN NOW Pillar 3. The Framework's adoption by major US federal agencies and its influence on private sector AI governance standards makes it a de facto compliance benchmark.

3.3

The Brazilian Context: A Unique Opportunity

Brazil occupies a structurally advantageous position in the global AI governance landscape for reasons that are underappreciated outside the country. As the world's sixth-largest

economy, the largest economy in Latin America, and a historically successful mediator in multilateral forums, Brazil has both the scale and the diplomatic track record to play an outsized role in the emerging international AI governance architecture.

Brazil's AI regulatory process — currently advancing through the National Congress with broad cross-party support — provides the domestic legitimacy for international leadership. ANPD's (National Data Protection Authority) framework under LGPD provides institutional infrastructure directly applicable to AI governance. And Brazil's leadership in the 2024 G20 — where AI governance was a priority agenda item — demonstrated the country's capacity and willingness to shape global norms on the issue.

The Brazilian advertising and communications sector, with its globally recognized creative capacity and its deep penetration of both traditional and digital media environments, is particularly well-positioned to lead on CONTAIN NOW Pillar 4 — epistemic containment. The skills that made Brazilian advertising the most decorated in the world — persuasion, narrative, cultural resonance — are precisely what an effective global containment movement requires to build the public will for governance action.

PART IV

The Manifesto: A Declaration and a Convocation

This is not a policy paper. It is a founding declaration. It is addressed to every person and institution capable of action.

WE DECLARE

We declare that artificial intelligence has crossed the threshold from technological development to civilizational transformation. This transition is not a future event. It is the defining condition of the present moment.

We declare that the pace of AI capability development has outrun — by a margin that is empirically documented and institutionally acknowledged — the pace of governance, alignment research, and social adaptation infrastructure. This gap is not a failure of intent. It is a failure of architecture. The architecture does not exist. Building it is the defining challenge of our time.

We declare that the framing of AI safety as antagonistic to AI progress is false, strategically counterproductive, and historically illiterate. Every transformative technology that created durable value — aviation, nuclear power, pharmaceuticals, financial markets — did so because governance frameworks were ultimately established. The technology that escapes governance does not create value; it creates liability, backlash, and, eventually, existential risk.

WE ASSERT

We assert that the most important economic opportunity of the 21st century is not AI capability development. It is AI containment infrastructure. The organizations — corporate, academic, governmental, civil society — that build the safety, alignment, interpretability, governance, and resilience infrastructure for AI will extract economic value proportional to the scale of the risk they are managing. That scale is civilizational.

We assert that the global AI safety research community — currently underfunded by two orders of magnitude relative to capability research — requires emergency resourcing. The gap is not a resource allocation problem awaiting normal budget cycles. It is a civilizational infrastructure problem requiring wartime mobilization logic.

We assert that international coordination on AI governance is not optional. It is structurally required. AI systems do not respect national borders. AI-enabled information operations do not respect democratic sovereignties. AI-driven economic disruption does not respect

existing social contracts. Governance frameworks that operate below the global level will be systematically arbitraged into irrelevance.

WE CONVOKE

We convoke governments to establish, fund, and empower national AI safety agencies with genuine regulatory authority; to collaborate on international governance instruments with binding force; and to price AI risk into their fiscal and procurement frameworks.

We convoke corporations to commit not merely to responsible AI principles — the world has enough principles — but to investment, governance structures, and operational practices that make responsibility measurable, auditable, and accountable. Safety must move from mission statement to balance sheet.

We convoke investors to recognize AI governance capacity as a material risk factor; to require ESG frameworks that specifically assess AI safety and alignment practices; and to fund the safety research and governance technology companies that the market currently under-rewards.

We convoke researchers to cross disciplinary boundaries; to prioritize publication, knowledge transfer, and policy translation alongside academic production; and to treat AI safety as what it is — the most important applied research challenge in the history of science.

We convoke citizens to demand accountability; to develop AI literacy as a civic competency; and to recognize that the governance of AI is not a technical problem to be delegated to experts. It is a political problem to be decided democratically. Democracy cannot survive AI-enabled epistemic disruption unless citizens understand what is happening to their information environment.

We convoke the AI industry itself — its engineers, researchers, executives, and investors — to act on what they already know. The public statements of the field's founders confirm awareness of the risk. Awareness without action is not responsibility. It is complicity.

"The time for caution about acting is over.
The time for urgency about not acting has arrived."

— CONTAIN NOW, 2025

PART V

Operational Roadmap: From Manifesto to Movement

5.1

Phase Architecture

PHASE 0 — IGNITION (2025)

- Publication and global distribution of this founding manifesto
- Establishment of CONTAIN NOW founding council with representation from science, industry, civil society, and government across minimum five continents
- Launch of CONTAIN NOW Index: a public-facing tracker of corporate AI governance commitments and verified compliance
- First CONTAIN NOW Summit — convening 500 organizations across 50 countries
- Partnership with at least two major multilateral institutions (UN, OECD, G20) for formal integration of CONTAIN NOW framework

PHASE 1 — ARCHITECTURE (2026–2027)

- Establishment of CONTAIN NOW Certification Standard — the first globally recognized voluntary certification for AI safety practices, modeled on ISO standards
- Launch of CONTAIN NOW Venture Fund: dedicated investment vehicle for AI safety and governance technology companies
- Pilot of AI Liability Insurance Framework in collaboration with major reinsurers
- Launch of CONTAIN NOW Academy: global curriculum for AI literacy and governance, freely available in all UN languages
- First Annual CONTAIN NOW Report: state of AI governance globally

PHASE 2 — INSTITUTIONALIZATION (2028–2030)

- CONTAIN NOW Certification adopted as a public procurement requirement in minimum 20 jurisdictions
- International AI Safety Agency proposal formally tabled at UN General Assembly
- CONTAIN NOW Venture Fund portfolio of 100+ AI safety companies across 30+ countries
- AI Literacy curriculum adopted in minimum 40 national educational systems
- First binding international instrument on AI governance substantially reflecting CONTAIN NOW framework

5.2

Metrics of Success

CONTAIN NOW explicitly rejects the metrics typically used by awareness movements — signatures, social media reach, media coverage — as insufficient and gameable. Success is measured by structural change in the following indicators:

Indicator	Description	Current	Target 2030
AI Safety Investment Ratio	Global ratio of safety/governance investment to capability investment	Current: <2%	**Target 2030: >15%**
Jurisdictional Coverage	% of global GDP covered by comprehensive AI governance frameworks	Current: ~28% (EU only)	**Target 2030: >70%**
Interpretability Capability	% of frontier AI deployments with certified interpretability tooling	Current: <5%	**Target 2030: >60%**
AI Literacy	% of population in OECD nations with assessed AI literacy	Current: ~8%	**Target 2030: >40%**
Incident Response	Time from AI incident detection to coordinated international response	Current: No mechanism	**Target 2030: <72 hours**

REFERENCES & SCIENTIFIC FOUNDATIONS

Selected Bibliography

• Acemoglu, D. & Johnson, S. (2023). Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity. PublicAffairs.

• Acemoglu, D. (2024). The Simple Macroeconomics of AI. NBER Working Paper 32487.

• Anthropic (2022). Constitutional AI: Harmlessness from AI Feedback. Bai, Y. et al. arXiv:2212.08073.

• Anthropic Interpretability Team (2023-2025). Towards Monosemanticity; Scaling Monosemanticity; On the Biology of a Large Language Model. transformer-circuits.pub.

• Bastani, H. et al. (2024). Generative AI Can Harm Learning. The Wharton School, University of Pennsylvania. SSRN Working Paper.

• Battiston, S. et al. (2016). Complexity theory and financial regulation. Science, 351(6275), 818-819.

• Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

• Brynjolfsson, E. et al. (2023). GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. NBER Working Paper 31161.

• Buldyrev, S.V. et al. (2010). Catastrophic cascade of failures in interdependent networks. Nature, 464, 1025-1028.

• European Parliament (2024). Regulation (EU) 2024/1689 — Artificial Intelligence Act. Official Journal of the European Union.

• Future of Humanity Institute, Oxford University (various, 2014-2024). Technical Reports on AI Safety and Existential Risk.

• Goldstein, J.A. et al. (2023). Generative Language Models and Automated Influence Operations. Georgetown CSET Report.

• Goldman Sachs Research (2024). AI Investment Forecast to Approach $200 Billion Globally by 2025. GS Global Investment Research.

• Hendrycks, D. et al. (2023). Aligning AI With Shared Human Values. ICML 2023.

• Krakovna, V. et al. (2020). Specification Gaming: The Flip Side of AI Ingenuity. DeepMind Blog / Nature Machine Intelligence.

• NIST (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. U.S. Department of Commerce.

• Olah, C. et al. (2022). Mechanistic Interpretability, Variables, and the Importance of Interpretable Bases. Distill / Transformer Circuits.

• Ord, T. (2020). The Precipice: Existential Risk and the Future of Humanity. Hachette Books.

• Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking/Penguin.

• Stanford HAI (2024). AI Index Report 2024. Stanford Institute for Human-Centered Artificial Intelligence.

• Taddeo, M. & Floridi, L. (2018). How AI can be a force for good. Science, 361(6404), 751-752.

• Tegmark, M. (2017). Life 3.0: Being Human in the Age of Artificial Intelligence. Knopf.

• UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence. UNESCO SHS/BIO/PI/2021/1.

• Wei, J. et al. (2022). Emergent Abilities of Large Language Models. Transactions on Machine Learning Research.

• World Economic Forum (2025). Future of Jobs Report 2025. WEF, Geneva.

• UK Department for Science, Innovation and Technology (2023). AI Safety Summit — Bletchley Declaration.

• UN Secretary-General's Advisory Body on AI (2024). Governing AI for Humanity: Interim Report.

• Various (2023). Statement on AI Risk. Center for AI Safety. Signed by 1,000+ AI researchers and executives.

CONTAIN NOW.
The most important movement you've never heard of.
Until now.

Join. Cite. Build. Fund. Govern. Contain.

CONTAIN NOW is an open movement. This manifesto is freely available for translation, adaptation, citation, and distribution. The only condition is attribution. The only objective is that it works.