

Corporate Responsibility and the Safe Development of AI

Professional ethics and public oversight in addressing the alignment problem

Brad Micallef • White paper • 10 September 2026

Executive summary

AI companies and the people directing them remain responsible for their choices even when government oversight is incomplete. Supporting useful AI is consistent with demanding restraint when its development or deployment creates serious risks that cannot be adequately controlled.

The claim that a company must accelerate because a competitor might act first identifies a real coordination problem. It does not establish that acceleration is safer, or excuse transferring poorly understood risks to the public. Regulation can address competitive pressures through enforceable common requirements. Professional ethics must guide the decisions being made while those requirements develop.

Five conclusions deserve immediate attention.

- **Responsibility already exists.** The UN business and human rights framework separates corporate responsibility from states' willingness or ability to act. Computing ethics codes require attention to harm, risk assessment, and the public interest. [1–3]
- **Restraint can support progress.** Asilomar's recombinant DNA deliberations illustrate scientists organizing precautions so research could proceed more responsibly. Thalidomide and Challenger demonstrate the importance of evidence, independent judgment, and effective escalation. [4–7]
- **Alignment is also a question of power and legitimacy.** Technical alignment concerns whether AI reliably behaves as intended. Social questions concern whose intentions count, who bears the risk, and who may authorize deployment. Researchers explicitly distinguish these issues. [8–10]
- **Practical action need not await a technical breakthrough.** Organizations can decline unsafe projects, limit autonomous authority, protect dissent, require independent approval, and choose alternatives supported by better safety evidence. Such choices can reduce exposure; they do not prove that a model is technically aligned.
- **Law and ethics reinforce one another.** Existing consumer protection law, California's frontier AI legislation, and the EU AI Act already impose relevant obligations within their respective scopes. Voluntary commitments are useful but insufficient as the sole basis for public protection. [15–19]

The immediate request: Developers, boards, investors, and professional leaders should adopt conditions for pausing or refusing unsafe work, document consequential decisions, and support independent public oversight.

Uncontrollable superintelligence is not an established or inevitable outcome. Responsibility requires proportionate precautions under uncertainty. [11] **What to build, empower, and release remains a human choice. If you direct or participate in those decisions, you are responsible for your actions. Competition and regulatory delay do not make that responsibility disappear.**

1 Purpose and scope

Author perspective

I am Brad Micallef. I am not an AI researcher. I am a technology and operations professional who has embraced advanced technology throughout my career and seen the benefits of using it responsibly.

My professional work includes energy technology, manufacturing, automation, controls, and the challenge of turning innovative systems into dependable operating capability. I serve as Executive Vice President, Manufacturing & Technology at Erthos and remain a founder and owner of Aderis Energy, with an ongoing business leadership and technical advisory role. These responsibilities put technology in the context of execution, reliability, and the people who depend on it. [23]

My concern comes from that practical perspective. Professionals routinely work within constraints, guidelines, and operating limits intended to prevent harm to themselves and those they serve. Exercising judgment about those limits is part of doing the job. Innovation does not remove that responsibility.

I do not claim that this experience qualifies me to resolve the technical alignment problem or predict when superintelligence might emerge. It does give me a basis to question the suggestion that competitive pressure or incomplete regulation makes professional restraint optional. Researchers bring essential technical expertise to that discussion. Practitioners and the public also have standing to ask what risks should be accepted on their behalf.

I offer this argument in my personal capacity. The companies named establish my professional background; this paper does not speak on their behalf.

The question this paper addresses

I support AI that helps people learn, work, create, and solve difficult problems. That support is the starting point of this paper. The purpose of technological development should be to improve human lives, with its benefits assessed alongside the risks and the distribution of both.

This paper argues that corporate responsibility and professional ethics apply before a government writes a detailed rule for every emerging capability. It addresses developers of advanced AI, the people who fund and direct their work, organizations deciding how to deploy it, and public officials designing oversight. It pays particular attention to recursive self-improvement, the alignment problem, and the possibility of losing meaningful human control.

The discussion examines an argument, not an allegation against every AI company: that competitive urgency makes restraint impractical and that governments must therefore assume responsibility for preventing dangerous outcomes. Requests for regulation can be sincere and constructive. The problem arises when those requests become an excuse for choices that decision makers still control.

The legal examples focus on the United States, California, and the European Union; international instruments supply broader ethical guidance. Legal status and sources were checked on 10 September 2026. This is a policy argument rather than a jurisdiction-specific compliance opinion or a finding that any company has violated the law. Historical comparisons illustrate decision-making principles; they do not establish the probability of an AI catastrophe.

The paper proceeds from the ethical argument to historical evidence, researchers' accounts of alignment, current standards and laws, and practical organizational commitments. The executive summary can be read independently. Section 9 provides a public statement suitable for adaptation.

2 Responsibility under competitive pressure

Ethics continues where detailed rules end

The absence of a specific prohibition is a weak basis for deciding whether an activity is acceptable. Rules necessarily operate at some level of generality, and organizations often know more about their own systems than outside observers do. That informational advantage strengthens the ethical case for disclosure and care.

The UN Guiding Principles on Business and Human Rights provide unusually direct support. Principle 11 and its commentary describe a corporate responsibility to respect human rights that exists independently of states' ability or willingness to meet their own obligations. OHCHR's public explanation reiterates that businesses are expected to respect rights even when states fail to do so. The framework also supports corporate due diligence and remedy. These principles are an authoritative normative framework, not a universal statute automatically creating the same liability everywhere. [1]

The ethical inference for AI is straightforward: a government's failure to govern adequately does not erase a developer's responsibility for foreseeable consequences within its influence. Equally, corporate responsibility does not relieve governments of their duty to establish and enforce public protections.

The strongest version of the race argument

A responsible developer may reasonably worry that slowing down will allow a less careful competitor to gain influence, set standards, or create dangerous capabilities. Delay can also postpone beneficial applications. These considerations deserve serious examination rather than dismissal.

But the conclusion that faster development makes society safer needs supporting evidence. A credible justification must explain why the particular acceleration reduces total risk, what alternatives were examined, how competitors are likely to respond, and what would make the company change course. A general promise to use future power responsibly cannot answer those questions.

If every participant treats another's potential recklessness as permission to relax its own safeguards, the argument provides no stopping point. This is precisely where common rules, independent verification, and lawful safety cooperation become valuable. Competitive pressure is a reason to build those arrangements and reassess dangerous choices; it is not a standing ethical exemption.

Benefit does not require zero risk

Many worthwhile activities involve risk. A defensible standard therefore asks whether a proposed activity has a legitimate benefit, whether a safer alternative can achieve it, whether precautions are credible, and whether the remaining burden is proportionate and fairly distributed.

The ethical objection becomes stronger when potential harm is severe, difficult to reverse, imposed on people who cannot meaningfully refuse it, and poorly understood by those making the decision. Conversely, uncertainty alone should not prohibit every experiment. Contained research with a clear learning purpose can be justified differently from releasing a system into critical infrastructure or granting it extensive independent authority.

The familiar observation that a person can create something dangerous without choosing to do so captures an important norm: capability does not create permission. An engineering comparison is more useful than a weapon analogy for AI, because AI has legitimate beneficial purposes. An engineer should not approve an inadequately supported design merely because a competitor might build it or an inspector has not intervened.

3 Historical lessons about restraint and accountability

Asilomar and recombinant DNA research

In the 1970s, researchers confronted uncertainty about the hazards of recombinant DNA experiments. David Baltimore's retrospective account describes scientists calling for a moratorium without having formal authority to impose one. Paul Berg describes the 1975 Asilomar conference as an effort to develop guidance that would allow the technology to be used safely. NIH issued recombinant DNA guidelines in 1976, replacing the conference recommendations for the research governed by those guidelines. [4–5]

The lesson is that people advancing a promising technology can organize restraint and help create the conditions for continued research. Professional initiative and institutional oversight can develop together.

The comparison has limits. A scientific meeting does not represent everyone exposed to risk, and contemporary AI competition has different commercial and international dynamics. Asilomar supports the feasibility of taking responsibility before comprehensive oversight; it does not show that voluntary agreements alone can govern the entire AI industry.

Thalidomide and the value of independent judgment

In 1960, FDA reviewer Frances Oldham Kelsey assessed an application for thalidomide. Despite pressure from the manufacturer and the drug's availability in other countries, she insisted on adequate scientific evidence of safety and did not approve the application. Severe birth defects were subsequently linked to the drug internationally. [6]

This is a case of professional judgment operating through existing regulatory authority, rather than self-regulation in a legal vacuum. Its relevance is the refusal to substitute commercial urgency or another jurisdiction's acceptance for adequate evidence.

For AI, the practical implication is to give reviewers the authority, time, and institutional protection to reject unsupported claims. A process that asks for expert review but punishes a negative finding undermines its own purpose. This comparison does not imply that every AI system should follow the same approval process as a medicine.

Challenger and the failure to act on warnings

The Rogers Commission's investigation of the 1986 Challenger accident found a flawed launch decision, failures of communication, conflict between engineering evidence and management judgment, and a structure that allowed safety concerns to bypass key managers. Its account includes continuing engineering opposition after the contractor's management changed its recommendation. [7]

The lesson extends beyond individual courage. An organization must preserve dissenting evidence and ensure that it reaches the people who can change a decision. A safety committee without access to decision makers, or a review process that records concerns without resolving them, is insufficient.

Challenger involved particular physical evidence and operating conditions. It cannot establish which AI capability is unsafe. It does establish why the presence of talented professionals and formal procedures is not, by itself, evidence that an organization will respond responsibly to warning signs.

Belmont and the distribution of benefits and burdens

The 1979 Belmont Report identifies respect for persons, beneficence, and justice as principles for research involving human subjects. It recognizes that learning may involve risk while requiring attention to minimizing harm, evaluating benefits, and distributing research burdens fairly. [12]

Applied by ethical analogy, these principles challenge the assumption that a socially beneficial objective justifies exposing any population to any degree of risk. AI deployment should consider people affected indirectly, including those who never agreed to use the product. Belmont is not a general AI law, and ordinary AI deployment is not automatically human-subject research. The useful connection is its disciplined treatment of consent, uncertainty, and unequal burdens.

4 The alignment problem and what researchers mean

Technical alignment

The alignment problem concerns whether AI behavior reliably serves intended human goals, interests, or values. These are not interchangeable targets. A system can follow an instruction while frustrating the person's actual purpose, and it can satisfy an operator while harming other people.

Richard Ngo, Lawrence Chan, and Sören Mindermann examine how advanced systems could learn behavior that appears acceptable in training but generalizes poorly, including deceptive behavior, misaligned internal goals, and strategies that increase their power. Their argument identifies possible failure mechanisms; it does not demonstrate that all present systems possess these properties or that catastrophe is inevitable. [8]

The Center for Human-Compatible Artificial Intelligence challenges the assumption that designers can simply specify a fixed objective and safely maximize it. CHAI describes research into systems that remain uncertain about human objectives and defer to people in light of that uncertainty. This is a technical research direction with a normative motivation, not a completed guarantee of control. [9]

Whose values and whose authority

Iason Gabriel distinguishes technical and normative aspects of alignment and emphasizes the differences among instructions, intentions, preferences, interests, and values. His argument gives central importance to fair principles that people with differing moral views can endorse. This supports treating alignment as a question of legitimate social choice as well as model behavior. [10]

A further sociotechnical critique by Adam Dahlgren Lindström and coauthors examines limitations of alignment through reinforcement learning from human feedback. The research questions whether such training can adequately represent complex human ethics and situates safety within institutions, processes, and technical systems. It supports examining who supplies feedback and how organizational choices shape the result. It does not establish that technical safeguards are unnecessary. [13]

For this paper, three questions help separate the issues. **Behavior:** will the system do what is intended under relevant conditions? **Legitimacy:** who is entitled to determine those intentions and resolve conflicts? **Institutional responsibility:** do incentives, approval procedures, and remedies protect the people affected? This is an organizing framework for the argument, not a claim that all researchers use the same taxonomy.

A technically obedient system can faithfully execute an unethical business objective. Conversely, a well-intentioned institution can deploy an unreliable system. Addressing either problem while ignoring the other leaves a serious gap.

Recursive self improvement and loss of control

Recursive self-improvement refers here to a feedback process in which AI contributes to producing improved AI that can then contribute further improvements. AI-assisted research is not, by definition, an autonomous runaway process. Capability growth, autonomy, access to resources, and the ability to circumvent oversight are distinct matters.

The International AI Safety Report 2026 describes evidence and uncertainty surrounding future loss-of-control risks. It finds insufficient evidence to reliably determine whether and how current capabilities and tendencies would scale into such outcomes. [11] This paper treats uncontrollable superintelligence as a possible future concern and assigns no numerical probability to it. That does not establish that every probability estimate is indefensible; it distinguishes uncertainty from inevitability.

Ethically, that uncertainty makes the design of decision points especially important. If an organization's planned development could outpace its ability to evaluate or interrupt the process, it should reconsider that plan before granting the relevant authority. Calling a system a tool does not establish its safety; the relevant questions concern what it can do, where it can act, and whether its effects remain manageable.

5 Social approaches to alignment that can begin now

Organizations do not need a new training algorithm to decide that some uses should not proceed. They can address the social dimensions of alignment and reduce exposure to technical failures through ordinary decisions about authority, professional standards, procurement, staffing, and accountability. The following are recommendations derived from the preceding argument, rather than claims that every measure is legally required.

Choose purposes through legitimate processes

Before approving a consequential application, specify the public or user benefit in concrete terms and identify whose interests might conflict with the purchaser's objectives. Include affected people, domain professionals, and independent reviewers in deciding acceptable uses. Give participation a defined effect on the decision, including a route to challenge it.

For example, an institution considering AI to allocate a scarce service should decide how fairness, eligibility, and appeals will work before asking a model to optimize allocation. Higher predictive accuracy cannot settle the underlying distributional judgment. Participation should include protections for minority rights; a simple majority preference is not a sufficient ethical test.

Withhold authority that cannot be responsibly supervised

Choose the least autonomous arrangement that achieves the legitimate purpose. An AI may draft an analysis without receiving authority to execute transactions, alter infrastructure, or make final decisions about people. Staff should retain independent information, enough time to review, and a real ability to refuse a recommendation.

This is a deployment choice, not an assertion that a nominal human approval step makes any system safe. If the volume, speed, or complexity of outputs exceeds meaningful review capacity, reduce the scope or decline the use. Existing access controls may help implement the decision; no claim is made that purely verbal rules can constrain an already uncontrollable system.

Make development choices reviewable

Require fresh approval when a project gains consequential new authority, changes its purpose, or uses AI to accelerate the creation of more capable successors. Separate authorization to investigate a capability from authorization to use it extensively or release it.

For a recursive development program, this can mean a board decision against allowing an automated process to authorize its own continuation. Require accountable people outside the immediate delivery chain to review each material expansion. If a stage cannot be evaluated before the next begins, the organization can decline that development pattern. This is a choice about whether to create the exposure, not a solution to every technical problem inside the model.

Align incentives with responsible decisions

Boards should ensure that safety leaders can escalate directly and suspend relevant work pending review. Promotion, compensation, and delivery targets should recognize identifying problems, improving reliability, and making justified decisions not to deploy. Record who approved a consequential decision, what uncertainty remained, and why alternatives were rejected.

Investors and customers can reinforce these practices through funding milestones and procurement terms that reward evidence of responsible conduct. Such commitments should be transparent enough to assess and enforce where appropriate. Responsibility should be strongest for those with the greatest control over resources and incentives; junior employees should not be expected to compensate for governance failures through personal sacrifice.

Protect disagreement and provide recourse

Offer confidential reporting, independent investigation, protection against retaliation, and a documented response to material safety concerns. Publish a clear route for affected people to obtain an explanation, seek correction, and reach a human decision maker. Where harm occurs, a review should lead to remedy and operational changes rather than merely a revised public statement.

Social norms become meaningful when they change behavior. Professional associations, employers, and purchasers can treat concealing serious risk or overriding sound safety objections without justification as failures of professional conduct. Procedures must also protect fair inquiry: a disputed judgment should be assessed on the evidence available at the time, not presumed dishonest merely because outcomes were poor.

Use restraint when uncertainty cannot be managed

Independent review should be able to produce four outcomes: approve the specified use, approve a narrower use, defer while evidence is obtained, or refuse the activity. Reviewers should state what new evidence could justify reconsideration. This avoids treating a pause as either meaningless delay or a permanent rejection of research.

These measures address alignment without waiting for a technological breakthrough by governing purposes and exposure. They do not guarantee reliable behavior, prevent every misuse, or solve international coordination. Where an intended use depends on a level of technical reliability or control that has not been demonstrated, the responsible social decision may be not to deploy it.

Illustrative applications

The following are hypothetical decision scenarios, not reports of particular deployments or evidence that these measures always succeed.

Proposed use	Decision available now	Remaining limitation
AI assists scientific literature review	Permit drafting with accountable expert verification	Verification takes expertise and time
AI recommends eligibility for a public service	Establish fair criteria and an independent appeal process before adoption	Accurate predictions do not establish fair policy
AI acts directly on critical infrastructure	Withhold execution authority unless supervision and safety evidence are adequate	Advisory output may still mislead operators

Proposed use	Decision available now	Remaining limitation
AI accelerates development of successor models	Require independent approval of material expansions and refuse unevaluable stages	A review process cannot guarantee model control
A serious internal concern remains unresolved	Defer the relevant release and investigate independently	The investigation must have authority and resources

6 Professional codes and international guidance

Computing ethics provides a direct professional foundation

The ACM Code of Ethics and Professional Conduct, adopted in 2018, identifies avoiding harm in Principle 1.2 and public-good leadership in Principle 3.1. Principle 2.5 calls for thorough assessment of systems and their risks, with special care for machine learning. Its guidance contemplates repeated reassessment or nondeployment where future risks cannot be reliably predicted. [2]

The joint IEEE-CS/ACM Software Engineering Code requires professionals to accept responsibility for their work, balance employer and client interests with the public good, approve software only on a well-founded safety basis, and disclose reasonably believed dangers to appropriate parties. These provisions appear in Principles 1.01–1.04. [3]

These are professional codes rather than generally applicable criminal or civil statutes. They nevertheless directly contradict the ethical proposition that a computing professional has no responsibility until a regulator supplies a specific instruction. Their practical force depends on adoption, professional enforcement, and organizational support.

International AI guidance supports continuing human responsibility

The OECD AI Principles, adopted in 2019 and updated in 2024, connect AI development with human rights, accountability, and management of risk throughout the lifecycle. Their safety principle calls for mechanisms to override, repair, or decommission systems that risk undue harm or exhibit undesired behavior. They are intergovernmental recommendations, not a universal corporate licensing law. [14]

UNESCO’s 2021 Recommendation on the Ethics of Artificial Intelligence centers human rights and dignity, with human oversight, impact assessment, and accountability. It provides a normative and policy framework rather than a directly enforceable global AI code. Its relevance is that responsibility extends to institutional choices surrounding a system, beyond its technical performance. [20]

NIST’s AI Risk Management Framework 1.0, published in 2023, offers the voluntary functions Govern, Map, Measure, and Manage. It can structure ownership of risk and continuing review. NIST’s current resource page notes that the framework is being revised; this paper refers specifically to version 1.0. Using a framework does not certify safety or establish that a particular risk is acceptable. [21]

AI developers have already recognized conditions for restraint

The 2024 Seoul Frontier AI Safety Commitments, whose government publication was updated in February 2025, include risk thresholds and processes for responding when those thresholds are reached. In extreme cases, signatories committed to forgo development or deployment when mitigations cannot keep risk below the defined thresholds. The commitments also address accountability and transparency. [15]

This is evidence that restraint is already part of the industry's stated vocabulary. It is not proof that every signatory has complied or that the commitments remain identically implemented. Any claim about a named company's current conduct requires a separate examination of its policies and decisions. The public can reasonably ask companies to explain how commitments affect actual development choices.

7 Legal examples and their limits

General consumer protection can reach new technologies

Section 5 of the US Federal Trade Commission Act, 15 U.S.C. § 45, prohibits unfair or deceptive acts or practices within the agency's jurisdiction. The statutory unfairness test concerns substantial actual or likely consumer injury that consumers cannot reasonably avoid and that is not outweighed by benefits to consumers or competition. Deception is a separate basis for enforcement. [\[16\]](#)

The implication is narrower than a general prohibition on dangerous AI research. Unsupported safety claims or harmful commercial practices may engage existing law if its requirements are met. Calling a product AI does not eliminate those requirements. Nor does a speculative societal risk automatically satisfy them. This illustrates how a general legal standard can operate without a rule written specifically for every technology.

California addresses frontier AI governance

California enacted SB 53, the Transparency in Frontier Artificial Intelligence Act, in September 2025. Business and Professions Code § 22757.12 requires covered large frontier developers to publish, implement, and comply with a frontier AI framework addressing catastrophic risk thresholds, mitigations, governance, and internal use. The legislation also provides for incident reporting and protects covered employees reporting specified dangers or violations. [\[17\]](#)

Coverage depends on statutory definitions, including training compute and revenue. The law does not require every AI developer to prove the absence of all risk. Its legislative findings connect safety with proportionate due care; this is not an unlimited standalone liability rule. [\[17\]](#) Its attention to internal use reinforces the ethical importance of decisions made before products reach customers.

The European Union imposes systemic risk duties

The EU AI Act, Regulation (EU) 2024/1689, establishes obligations for providers of general-purpose AI models and additional duties for those with systemic risk. Article 55 addresses evaluation, assessment and mitigation of systemic risks, serious-incident handling, and cybersecurity. The AI Office's explanatory Q&A describes certain duties concerning development of models intended for the EU market. It also acknowledges the Act's general exclusion of research, testing, or development before market placement or entry into service. The Q&A is explanatory, not an authoritative judicial interpretation or official Commission position. [\[18\]](#)

The general-purpose AI obligations began applying on 2 August 2025, with transition provisions for models already on the market. Commission guidance identifies 2 August 2027 for compliance by providers of models placed on the market before 2 August 2025. These dates concern general-purpose models and should not be generalized to every category of AI system. [\[19\]](#)

The relevant point is that identifying and managing systemic risk can be a provider obligation. Compliance is not a guarantee of harmlessness, and the law's scope and classifications must be established for any specific model. Voluntary implementation guidance should also be distinguished from the binding obligations it helps providers meet.

Broad duties can supplement detailed standards

Section 5(a)(1) of the US Occupational Safety and Health Act requires covered employers to provide a workplace free from recognized hazards causing or likely to cause death or serious physical harm. Its general duty structure illustrates that law can impose responsibility beyond a catalog of specific technical instructions. [22]

This is an analogy about legal design. The provision protects workers within its statutory scope; it does not create a general duty under that Act to prevent every societal AI risk. The distinction matters because an ethical argument loses credibility when it presents a useful analogy as an existing universal legal obligation.

Comparison of the principal instruments

Instrument	Type of authority	Relevance to this argument
UN Guiding Principles [1]	International normative framework	Corporate responsibility persists despite state failure
ACM and IEEE-CS/ACM codes [2–3]	Professional ethics	Public interest and reasoned safety judgment
OECD and UNESCO guidance [14, 20]	Intergovernmental recommendations	Human oversight and accountability
NIST AI RMF 1.0 [21]	Voluntary risk framework	Assigning and managing risk responsibilities
Seoul commitments [15]	Voluntary corporate commitments	Explicit thresholds and restraint
FTC Act § 5 [16]	US federal law within its jurisdiction	Unfair or deceptive commercial conduct
California SB 53 [17]	State law for defined covered actors	Frontier risk governance and reporting
EU AI Act [18–19]	EU regulation within its scope	Duties concerning systemic risk
OSH Act § 5(a)(1) [22]	US workplace safety law	Analogy for a general duty of care

8 A practical standard for responsible development

The decision that should precede consequential expansion

Before substantially expanding capability, autonomy, access, or deployment, the accountable organization should produce a concise decision record addressing six questions. These are proposed governance requirements, not a claim of universal existing law.

1. **Purpose:** What benefit is sought, for whom, and why is this level of capability necessary?
2. **Exposure:** Who could be harmed, including nonusers, and how severe or irreversible could the harm be?

3. **Evidence:** What supports the safety judgment, what remains unknown, and what evidence would change the decision?
4. **Alternatives:** Could a narrower system, reduced authority, or a different workflow achieve the benefit more safely?
5. **Authority:** Who independently reviews the decision, who can stop the activity, and how can affected people challenge it?
6. **Conditions:** What events require reassessment, suspension, notification, or remedy?

Specific warning conditions may include a credible failure of oversight, a new capability outside the assessed scope, an unresolved independent finding of serious risk, or inability to review important actions before they take effect. Organizations should define such conditions before commercial deadlines dominate the discussion. A threshold that changes whenever it becomes inconvenient has little protective value.

What different participants should do

Boards and executives should approve risk tolerances, fund independent review, and make clear that reviewers can delay consequential work. They should require an explanation of how public safety commitments have affected actual decisions. Researchers and engineers should document material uncertainties, communicate concerns, and decline professional endorsement when the evidence does not support it.

Investors should avoid conditioning success exclusively on speed or benchmark performance. Customers should ask what authority the system receives, who remains accountable, and what happens when the system is wrong. Professional associations should translate codes into practical guidance and credible processes for reviewing alleged misconduct.

Governments should establish enforceable minimum requirements, support independent evaluation, protect reporting, and build international coordination. They should include affected communities and avoid allowing the largest developers to define standards solely around their own resources or business models. Proportionate requirements can focus scrutiny on consequential uses and serious risks while preserving access to beneficial innovation.

Answering the principal objections

Restraint may hand an advantage to less responsible actors. That risk should be assessed. It supports enforceable common rules and carefully designed cooperation. A proposed acceleration should still demonstrate why it is safer than alternatives rather than presume that the identity of its developer settles the matter.

No one can prove an advanced system perfectly safe. Perfect safety is not the standard proposed here. The question is whether evidence and safeguards justify the particular exposure. Where a decision could create severe, irreversible effects, a lack of credible control can justify refusal even without certainty that harm will occur.

Ethics is subjective. People disagree, including about acceptable risk. That is why the paper emphasizes rights, transparent reasons, participation, independent review, and appeal. Leaving the decision entirely to a company does not remove value judgments; it concentrates them.

Social governance cannot control a superintelligence. It may not. Its preventive role is to govern whether and under what conditions such a system is developed, empowered, or released. Ethical restraint is relevant precisely when confidence in technical control is inadequate. These measures should accompany technical research without assuming that a future solution excuses present exposure.

Regulation can protect incumbents and slow useful work. Poorly designed rules can do so. The response is risk-proportionate oversight, accessible evaluation, transparent standards, and attention to competition.

Voluntary action also needs scrutiny because private safety claims can serve commercial interests. Public accountability remains essential in either case.

9 Public statement

I support the development of AI that improves people's lives. Its potential to advance knowledge, expand opportunity, and help solve difficult problems is worth pursuing. That promise creates a responsibility to exercise judgment about what we build, how much authority we give it, and when we should stop.

Government regulation is necessary. Its absence does not suspend corporate responsibility or professional ethics. Companies and individuals remain accountable for decisions within their control and for the serious risks they knowingly create or fail to examine responsibly.

The argument that we must move faster because someone else will cannot, by itself, justify exposing society to harm. Every competitor can make that claim. A race does not become responsible because each participant believes it should win.

AI researchers describe an alignment problem: ensuring that powerful systems reliably serve human purposes and values. Part of that challenge is technical. But deciding whose purposes count, what risks are acceptable, and whether a system should receive consequential authority is also a social and ethical responsibility.

We do not need to solve every technical alignment question before making responsible choices. We can refuse inadequately supported deployments, protect professionals who raise concerns, require independent scrutiny, and preserve meaningful human authority. Where reliable control cannot be demonstrated, restraint is a decision available now.

The possibility of AI accelerating its own development makes those decisions more urgent. It does not make catastrophic loss of control inevitable, and it does not make the pursuit of every capability necessary. Useful progress should be judged by the human benefit it delivers and the conditions under which it is achieved.

I call on developers, executives, boards, and investors to make their safety commitments consequential: establish clear conditions for delay or refusal, disclose serious risks responsibly, support enforceable oversight, and accept accountability for the choices they make while public institutions catch up.

Supporting AI includes insisting that its development respects human agency, public safety, and the people who bear its risks.

The path ahead is shaped by human choices. People choose what to fund, what to build, what authority to grant, and what to release. No individual controls the entire industry, but each remains responsible for their own actions, and those with greater authority bear greater responsibility.

If you are involved in these decisions, the question is what you will choose to do with the knowledge and authority you have. Competitive pressure does not make that choice inevitable. Regulatory delay does not make it someone else's choice. **Ultimately, you are responsible for your actions.**

References

Sources accessed or checked on 10 September 2026. Numbered citations identify the source for factual claims. Recommendations and applications to AI are the argument of this paper; the inclusion of a source does not imply its authors endorse the entire paper.

- [1] United Nations, Office of the High Commissioner for Human Rights. Guiding Principles on Business and Human Rights (2011), especially Principles 11, 17, and 22; OHCHR, Business and Human Rights, explanation of responsibility despite state failure. [Open source](#) | [Open source](#)
- [2] Association for Computing Machinery. ACM Code of Ethics and Professional Conduct (2018), Principles 1.2, 2.5, and 3.1. [Open source](#)
- [3] IEEE Computer Society and ACM. Software Engineering Code of Ethics and Professional Practice, version 5.2, Principles 1.01–1.04. [Open source](#)
- [4] NobelPrize.org. David Baltimore interview (2001), discussion of the recombinant DNA moratorium; Paul Berg interview (2001), discussion of Asilomar. Participant accounts. [Open source](#) | [Open source](#)
- [5] National Institutes of Health. NIH Guide for Grants and Contracts, vol. 5, no. 20 (30 November 1976), notice of guidelines issued 23 June 1976 and institutional review procedures. [Open source](#)
- [6] US Food and Drug Administration. Frances Oldham Kelsey Medical Reviewer Famous for Averting a Public Health Tragedy. Agency historical account. [Open source](#)
- [7] Presidential Commission on the Space Shuttle Challenger Accident. Report (1986), volume I, chapter V, The Contributing Cause of the Accident. NASA historical archive. [Open source](#)
- [8] Richard Ngo, Lawrence Chan, and Sören Mindermann. The Alignment Problem from a Deep Learning Perspective. ICLR 2024; arXiv version 8 (4 May 2025). [Open source](#)
- [9] Center for Human-Compatible Artificial Intelligence. Research overview, discussion of uncertainty about objectives and deference to humans. [Open source](#)
- [10] Iason Gabriel. Artificial Intelligence, Values, and Alignment. Minds and Machines (2020). [Open source](#) | [Open source](#)
- [11] Yoshua Bengio and coauthors. International AI Safety Report 2026 (3 February 2026), discussions of AI research and development and loss-of-control risks. [Open source](#)
- [12] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The Belmont Report (1979), Part B and risk–benefit assessment. US Department of Health and Human Services. [Open source](#)
- [13] Adam Dahlgren Lindström and coauthors. Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through Reinforcement Learning from Human Feedback. Ethics and Information Technology 27, article 28 (4 June 2025). [Open source](#)
- [14] Organisation for Economic Co-operation and Development. OECD AI Principles (2019, updated 2024), robustness, security and safety; accountability. [Open source](#)
- [15] UK Department for Science, Innovation and Technology. Frontier AI Safety Commitments, AI Seoul Summit 2024 (updated 7 February 2025), Outcomes 1–3, especially paragraph IV. [Open source](#)
- [16] Federal Trade Commission Act § 5, 15 U.S.C. § 45(a), (n). FTC explanation in A Brief Overview of the Federal Trade Commission’s Investigative, Law Enforcement, and Rulemaking Authority. [Open source](#)

- [17] California SB 53 (2025), chapter 138, approved 29 September 2025. Business and Professions Code §§ 22757.10–22757.16 and Labor Code provisions beginning at § 1107; especially § 22757.12 and legislative finding 1(p). Enacted text. [Open source](#)
- [18] Regulation (EU) 2024/1689, especially Articles 2(8) and 51–55. AI Office, General-Purpose AI Models in the AI Act Questions and Answers, explanatory discussion of development-stage obligations and exclusions; the page disclaims official Commission status. [Open source](#) | [Open source](#)
- [19] European Commission. Guidelines on Obligations for General-Purpose AI Providers; implementation dates and transition arrangements. [Open source](#)
- [20] UNESCO. Recommendation on the Ethics of Artificial Intelligence (adopted 2021); official publication overview and explanation of oversight and accountability. [Open source](#) | [Open source](#)
- [21] Elham Tabassi. Artificial Intelligence Risk Management Framework AI RMF 1.0. NIST AI 100-1 (26 January 2023). NIST resource page also identifies ongoing revision. [Open source](#) | [Open source](#)
- [22] Occupational Safety and Health Act of 1970, § 5(a)(1), 29 U.S.C. § 654(a)(1). OSHA statutory text. [Open source](#)
- [23] Brad Micallef. Public LinkedIn profile and displayed announcement of appointment at Erthos and continuing role at Aderis Energy. Professional biography is self-reported. [Open source](#)