WHO GOVERNS "SUPER INTELLIGENCE"?

A Christian Reflection of the White House Self-Regulation Accord, Technological Power, and Human Flourishing

Hoong Chuin LAU

3 October 2026

AI Assistance Disclosure: Generative AI tools were used in the research, drafting, synthesis, and editorial development of this paper. The author determined the paper's arguments, theological framing, source selection, revisions, and conclusions, and reviewed the final text for accuracy and interpretation. Responsibility for the views expressed and any remaining errors rests with the author.

Disclaimer: The views expressed are my own and do not represent my university or employer.

Contents

Executive Summary
1. The September 29 Turn: From AI to “Super Intelligence”
2. What the White House Accord Actually Does
3. The Strongest Case for Self-Regulation
4. The Empirical Problem: Voluntary Commitments Have a Verification Gap
5. Moral Hazard and the Structure of Competition
6. Who Regulates the Regulators—and Who Audits the Auditors?
7. Democratic Legitimacy: Expertise Is Not the Same as Authority
8. Intelligence Is Not Wisdom
9. Magnifica Humanitas: Human Primacy and the Governance of AI
10. Human Flourishing as a Governance Criterion
11. Governance Alternatives and Their Trade-offs
12. A Framework for Evaluating the White House Accord
13. Conclusion: The Governance Problem Behind the Intelligence Problem
References

Executive Summary

On 29 September 2026, U.S. President Donald J. Trump issued Executive Order 14434, Inaugurating the Era of Super Intelligence. The order directs executive departments and agencies, to the maximum extent permitted by law, to replace the terms "Artificial Intelligence" and "AI" with "Super Intelligence" and "SI" in official non-statutory communications. For purposes of the order, however, "Super Intelligence" initially means the same technologies encompassed by the existing statutory definition of artificial intelligence. The order also directs the Assistant to the President for Science and Technology to propose legislative language for a federal definition of "Super Intelligence."

On the same day, the President and leaders of major frontier-technology companies signed the "**White House Accord on Super Intelligence**: Joint Commitment on Frontier Responsibilities". The accord states that each company is responsible for developing its own technology safely and identifies four layers of control: internal safeguards, an internal oversight team, an independent external auditor or evaluator, and an independent board committee. The accord is voluntary and states that its provisions may later be codified into law or regulation.

These two actions create a significant governance question. The federal government has declared AI as entering an era of "Super Intelligence," while the principal mechanism for governing the associated frontier risks remains corporate self-regulation backed by voluntary auditing and board oversight. The central issue is therefore not whether self-regulation has value. It plainly can: frontier developers possess technical knowledge, operational visibility, specialist personnel, and the ability to modify systems faster than public institutions can normally legislate. The more difficult question is whether self-regulation is institutionally sufficient when the technologies at issue are simultaneously represented as transformative enough to alter the vocabulary of government and potentially reshape economic, scientific, political, and social life.

This paper examines that question through three tests.

First, **epistemic adequacy**: Can the governance structure discover, measure, and respond to frontier risks effectively? Frontier AI risks evolve rapidly, and the companies developing the systems possess much of the relevant expertise. This supports a substantial role for internal governance. Yet empirical research on voluntary commitments and frontier safety frameworks identifies important weaknesses in public verification, measurable risk tolerances, pause thresholds, and identification of unknown risks [Wan25, Ste25].

Second, **institutional accountability**: Can external actors independently verify corporate claims and impose consequences when safeguards fail? Research on frontier AI governance generally treats industry self-regulation as a useful first layer rather than a complete governance regime. It emphasizes independent access, reporting, standard setting, regulatory capacity, and some mechanism for ensuring compliance [And23, Sch24, Bru26]. Auditing can improve accountability, but only where auditors have meaningful access, adequate standards, professional independence, and a clear relationship to public authority [Cos22, Bru26].

Third, **normative legitimacy**: Who has the authority to decide which risks are acceptable and what purposes advanced AI should serve? Decisions about cybersecurity, biosecurity, or model reliability contain technical components, but broader questions concerning employment, surveillance, inequality, autonomy, warfare, education, social dependency, and the distribution of technological power are not merely engineering questions. They involve contested judgments about the common good. Scholarship on democratic AI governance therefore warns against allowing technical expertise to substitute for public deliberation [Coe24].

A Christian ethical analysis adds a further distinction: *intelligence is not identical to wisdom*, and *capability is not identical to human worth*. Antiqua et Nova and Pope Leo XIV's 2026 encyclical Magnifica Humanitas argue, in complementary ways, that human intelligence involves embodied, relational, moral, and purposive dimensions that cannot be reduced to computational performance. Human dignity does not depend on outperforming machines. From this perspective, the relevant ethical question is not simply whether increasingly capable systems can be made safe, but whether their development and governance preserve responsible human agency, subsidiarity, solidarity, stewardship, participation, and the common good.

Scripture deepens this moral framework. Genesis 1:27 grounds human dignity in the image of God; Psalm 1:1–3 & John 10:10 portray flourishing as abundant and fruitful life; Micah 6:8 joins justice, mercy, and humility; Proverbs 4:7 and James 3:17 distinguish wisdom from knowledge or power; Genesis 2:15 and Luke 12:48 frame human capability as stewardship for which responsibility is required; and Mark 8:36 warns against treating worldly gain as the highest good. These texts provide moral ends and virtues by which technological power can be judged.

More directly for the governance question examined here, Magnifica Humanitas places AI within Catholic Social Doctrine and insists on the primacy of the human person. It calls for human conscience and freedom to determine the use and limits of technology; robust legal frameworks and independent oversight; participation and subsidiarity where AI affects public goods and fundamental rights; and scrutiny of the concentration of data, infrastructure, and computational power. Its culminating test is whether technological development makes human life more human and more worthy of the person.

The paper therefore treats the White House initiative neither as proof that self-regulation is inherently irresponsible nor as evidence that formal regulation is automatically superior. Rather, it uses the initiative as a case study in a larger institutional problem: as the claimed power of a technology increases, the distinction between technical expertise and legitimate authority becomes more consequential.

> *The greater the societal consequences attributed to an intelligent system, the stronger the case for distinguishing corporate responsibility for building it from public legitimacy in governing the risks it creates.*

The long-term challenge is consequently not only the governance of "superintelligence." It is the governance of human institutions that possess, deploy, and benefit from increasingly powerful forms of computation.

1. The September 29 Turn: From AI to "Super Intelligence"

President Trump's Executive Order 14434 does something unusual in technology policy. It does not merely introduce a program, fund research, or regulate a particular class of systems. It changes the vocabulary through which the executive branch is instructed to describe the field.

The order states that today's frontier systems increasingly "amplify human ingenuity" and unlock new forms of creativity, and that their capabilities represent "not merely artificial intelligence, but a new era of Super Intelligence." It directs agencies to substitute "Super Intelligence" and "SI" for "Artificial Intelligence" and "AI" in non-statutory documents and communications. Yet the order's operative definition does not establish a new technical threshold. Until superseded, "SI" is defined by reference to the existing statutory definition of artificial intelligence.

This distinction matters. In much of the technical and philosophical literature, "superintelligence" normally describes a hypothetical level of general cognitive performance substantially beyond human capability across many domains. Executive Order 14434 instead uses "Super Intelligence" as an administrative and rhetorical category encompassing technologies already legally classified as AI.

The redefinition therefore operates at two levels.

At the legal-administrative level, the immediate change is principally terminological. Existing AI systems are relabeled "SI" within executive communications, while the underlying statutory definition remains temporarily unchanged.

At the political-symbolic level, the shift is more ambitious. It presents frontier computation as a civilizational technology whose promise is better captured by the word "super" than by the word "artificial." In public remarks the same day, President Trump described the technology as "extreme," "superior," and "supreme" intelligence, said it could be bigger than the Industrial Revolution, and framed international leadership as a competitive race in which the United States should maintain its lead.

This language does not by itself determine a regulatory outcome. It does, however, establish the policy context in which the accompanying accord should be interpreted. The technology is simultaneously characterized as unusually powerful, rapidly advancing, economically strategic, and difficult to constrain without risking national competitive advantage.

That combination produces a recurring problem in technology governance: the same characteristics used to justify urgency in development can also make oversight appear too slow, too rigid, or too dangerous to competitiveness.

The resulting question is not merely semantic. If ordinary AI and hypothetical superintelligence are placed under the same rhetorical label, the term can obscure important distinctions among present capabilities, frontier capabilities, AGI scenarios, and genuine superintelligence. Conversely, the terminology may also signal that the administration intends to treat frontier AI as a strategic technological platform rather than a collection of discrete software tools.

A rigorous analysis therefore needs to separate three propositions:

- Current frontier systems are economically and scientifically powerful.
- Future systems may become substantially more capable.
- Current systems are already "superintelligent" in the conventional technical sense.

The executive order clearly embraces the first and anticipates the second. It does not technically establish the third, because its present definition deliberately inherits the statutory definition of AI.

That conceptual precision is important for governance. Regulation calibrated to current models is not necessarily adequate for future systems with substantially greater autonomy or strategic capability. Conversely, rules designed for hypothetical superintelligence may be disproportionate when applied to ordinary low-risk AI systems. Terminology should therefore be treated as consequential when it collapses governance categories that may require different levels of scrutiny.

Primary sources: Executive Order 14434 and White House Fact Sheet

2. What the White House Accord Actually Does

The White House Accord on Super Intelligence is compact. Its brevity is one of its strengths and one of its limitations.

The accord begins with a clear assignment of responsibility: "every company is responsible for developing its own technology safely." It then describes four layers of controls.

- Internal controls monitoring model capabilities and alignment during training and deployment, particularly around cybersecurity, biosecurity, chemical threats, and unauthorized access to technical systems.
- An internal team empowered to ensure that controls, monitoring, and detection work as intended and that identified problems are remediated.
- An independent external auditor or evaluator to assess whether these mechanisms operate as intended.
- An independent board committee to receive reports from internal and external assurance functions and oversee remediation.

The accord also says participating companies will meet regularly to establish standards and best practices. It explicitly acknowledges that the controls might later be codified in law or regulation.

This is not an absence of governance. It resembles a recognizable layered assurance model used in financial services, cybersecurity, safety engineering, and corporate risk management, augmented by board oversight. Internal controls can respond rapidly to technical change. Dedicated safety teams create organizational accountability. External evaluation can reduce self-certification. Board committees can elevate safety issues above ordinary product management.

For that reason, the serious critique of the accord is not that its four layers are meaningless. Each can be valuable. The analytical issue is what the document leaves unspecified.

Omissions: The accord does not itself define measurable risk tolerances; nor does it specify the capability or risk threshold that requires a company to pause training or deployment, nor identify accreditation requirements for external auditors, nor establish what technical or organizational access an auditor must receive. It does not require publication of audit findings, nor define how conflicts of interest between auditors and clients will be managed. It does not create statutory penalties for non-compliance, nor specify a public authority that can independently prohibit deployment when a company and its auditor disagree about the adequacy of safeguards.

These omissions do not prove that the framework will fail. Many operational details could be developed through standards and company practice. But they matter because the efficacy of a voluntary framework depends heavily on the credibility of precisely these institutional details.

Research on algorithmic auditing is instructive. Costanza-Chock et al. found broad support among auditing practitioners for clearer standards, independent scrutiny, disclosure, affected-stakeholder participation, and mechanisms that connect findings to actual remediation [Cos22]. More recent work on frontier AI auditing argues that meaningful assurance requires deep and secure access not only to public model interfaces but to internal safety cases, incident logs, relevant organizational processes, and in higher-assurance settings even model internals and compute records [Bru26].

> *The phrase "independent external auditor" is not self-executing. Independence has to be designed.*

The same is true of board oversight. A committee can be independent in a corporate-governance sense while still operating within an institution whose fiduciary, competitive, and strategic interests are implicated by the technology. Board oversight is therefore an important form of corporate accountability, but it is conceptually distinct from public accountability.

Primary source: White House Accord on Super Intelligence

3. The Strongest Case for Self-Regulation

A credible critique should first state the case for self-regulation in its strongest form.

3.1 Frontier AI changes faster than legislation

Technical capabilities, model architectures, deployment patterns, and evaluation methods can change in months. Detailed statutory rules can therefore become obsolete before they are implemented. A rule written around a particular benchmark, parameter count, or architecture may quickly become irrelevant or create incentives to design around the threshold.

Schuett et al. therefore argue that frontier AI regulation cannot sensibly be reduced to a choice between detailed rules and broad principles. In an immature and rapidly changing field, higher-level principles can preserve adaptability while developers innovate on the technical means of compliance [Sch24].

3.2 The companies possess much of the relevant knowledge

Frontier developers know their training systems, architectures, internal evaluations, model behavior, deployment stack, and security practices in ways external institutions often do not. Much of the world's specialized frontier-AI talent is concentrated in the private sector.

This produces a genuine information problem for government. A regulator cannot simply write technically perfect rules from outside the system. Effective governance therefore requires the cooperation of the developers themselves [And23, Sch24].

3.3 Internal controls can act before public law

When a novel capability emerges during training, the first institution capable of detecting and responding to it is normally the developer. Internal red teams, model evaluations, secure development procedures, staged deployment, incident monitoring, and release gates can operate at machine-development speed.

Voluntary commitments can also establish norms before legislation catches up. Anderljung et al. describe self-regulation as a potentially valuable first step in building standards, assurance capacity, and organizational safety practices [And23].

3.4 Overly prescriptive regulation can create its own risks

Rigid regulation may entrench incumbents, impose disproportionate compliance costs on smaller firms and academic research, or lock immature safety practices into law. It can also create "checkbox compliance," where meeting a formal requirement becomes a substitute for genuine risk reduction [Sch24].

The case for self-regulation is therefore not simply ideological. It rests on real features of frontier technology: speed, uncertainty, expertise concentration, and the need for adaptive controls.

The central question is whether these strengths justify self-regulation as the final layer of authority, or instead justify self-regulation as one component of a larger governance structure.

4. The Empirical Problem: Voluntary Commitments Have a Verification Gap

The new accord does not emerge in an institutional vacuum. The United States has already experimented with voluntary White House commitments from AI developers.

Wang, Huang, Klyman, and Bommasani examined public evidence concerning sixteen companies' implementation of the 2023 White House voluntary AI commitments [Wan25]. Their study is important but must be interpreted carefully: it assessed publicly available evidence of implementation, not every confidential practice a company may have undertaken.

Within that limitation, the findings are significant.

Across the companies studied, the mean public-evidence score was roughly half of the available indicators. Performance varied sharply among firms. Some areas, particularly model-weight security, systematic information sharing, comprehensive red teaming, and third-party reporting channels, showed substantial gaps. The authors also noted a fundamental accountability problem: except where commitments explicitly required public reporting, a company could potentially satisfy a commitment without providing evidence that allowed the public to verify it.

> *A commitment can exist without a verification architecture.*

This does not mean voluntary commitments are useless. The better-performing companies demonstrated that voluntary action can generate substantive practices. The study instead suggests that the public credibility of such commitments depends on specification, disclosure, and verifiability.

The newer frontier safety frameworks raise a related concern. Stelling et al. evaluated twelve company frameworks across risk identification, risk analysis, risk treatment, and governance [Ste25]. They found substantial variation and broad gaps. Particularly relevant were the absence of measurable risk tolerances, limited commitments to clear development or deployment pauses, weak processes for discovering previously unknown risks, and considerable managerial discretion in deciding how safeguards would be applied.

The implication is not that companies lack safety expertise. It is that technical expertise does not automatically produce externally legible accountability.

This distinction becomes more important as corporate safety frameworks move from internal management documents to quasi-public governance instruments. When governments or the public rely on those frameworks, their function changes. They are no longer merely tools for internal learning; they become evidence that risks are being governed.

That shift creates a higher burden of clarity.

5. Moral Hazard and the Structure of Competition

The most important ethical criticism of self-regulation does not require assuming bad faith by corporate leaders.

It arises from incentives.

A frontier AI company that discovers a serious risk may face several competing obligations and pressures: user safety, shareholder expectations, employee concerns, strategic partnerships, national-security narratives, capital markets, and competitive pressure from rival laboratories.

Suppose one company concludes that a capability threshold justifies delaying deployment. If competitors continue, restraint may impose substantial commercial and strategic costs on the cautious actor. If policymakers simultaneously describe international AI development as a race with a winner and loser, those pressures intensify.

This is a classic collective-action problem.

Individual firms may rationally prefer an industry in which everyone follows strong safety practices while simultaneously having incentives to move faster if they believe others will do so. The result can resemble a race dynamic even when no participant actively desires unsafe outcomes.

Schuett et al. therefore identify misaligned incentives, negative externalities, and competitive dynamics as reasons that internal expertise and external oversight are complementary [Sch24]. Anderljung et al. make a similar argument: industry self-regulation can seed good practice, but some mechanism for ensuring compliance is needed when severe public risks are possible [And23].

> *A governance system is fragile when socially desirable restraint imposes private costs on the actor expected to exercise it.*

This is why moral character alone cannot carry the entire burden of governance.

Corporate virtue matters. Leadership judgment matters. Safety culture matters. But durable institutions are normally designed on the assumption that incentives sometimes diverge from the public interest.

This is equally true of government regulation, which can suffer from political incentives, bureaucratic failure, and capture. The appropriate comparison is therefore not "virtuous companies" versus "virtuous government." It is among institutional arrangements with different failure modes.

6. Who Regulates the Regulators and Who Audits the Auditors?

The accord's strongest answer to the self-regulation problem is external auditing. However, that answer deserves serious attention.

A rigorous external audit can reduce information asymmetry between developers and the public. It can test whether internal controls exist in practice, reproduce internal evaluations, examine decision processes, identify gaps, and provide assurance that company claims are not purely self-certified.

But frontier AI auditing is still an emerging profession.

Brundage and a large group of collaborators argue that meaningful frontier auditing requires "deep, secure access" to non-public information and must examine the organization, not merely a public model endpoint [Bru26]. Depending on the level of assurance sought, auditors may need access to safety cases, staff interviews, internal evaluations, incident logs, compute records, model variants, governance processes, and technical systems.

They also emphasize independence safeguards: avoiding financial dependence on a single client, separating auditing from consulting, managing conflicts of interest, accrediting auditors, and giving oversight institutions the capacity to inspect audit quality.

This matters because a nominally external auditor can still be structurally dependent on the company paying for the audit.

Financial auditing confronted the same issue long before AI. An auditor can be legally separate from a client and still face commercial incentives not to alienate that client. The institutional challenge is therefore not only independence in name but independence in incentives, access, methods, and reporting.

Costanza-Chock et al. put the problem succinctly in the title of their study: Who Audits the Auditors? Their field scan found that algorithmic auditing can become symbolic when standards are unclear, findings remain confidential, remediation is optional, or auditors themselves lack competence and accountability [Cos22].

The White House accord's external-audit requirement may therefore represent a meaningful advance if it stimulates a mature assurance ecosystem. But its value depends on unresolved design questions:

- What constitutes independence?
- Who selects and pays the auditor?
- What information must the company provide?
- What standards does the auditor apply?
- Can the auditor inspect safety-critical non-public systems?
- Must the auditor disclose material findings?
- What happens when management rejects the auditor's conclusion?
- Can a regulator or other public body review the audit?
- Who accredits or disciplines incompetent auditors?

These are not secondary implementation details. They determine whether auditing is a verification mechanism or a reputational signal.

7. Democratic Legitimacy: Expertise Is Not the Same as Authority

The governance debate becomes more difficult when it moves from technical hazards to social choices.

Technical questions include:

- Does a model materially increase cyber-offensive capability?
- Can safeguards prevent unauthorized access?
- Can an evaluation reliably detect a dangerous biological capability?
- How secure are model weights?

These questions require specialist expertise.

But AI governance also entails questions such as:

- What level of unemployment or occupational displacement is socially acceptable?
- How much surveillance should institutions be permitted to use?
- When should automated decisions be contestable?
- Which domains should preserve non-automated human judgment?
- How should the gains from automation be distributed?
- When does reliance on AI undermine meaningful human agency?
- What degree of catastrophic risk may society legitimately accept in exchange for faster technological progress?

These are not purely technical questions. They involve competing values, distributional consequences, rights, and conceptions of the good society.

Coeckelbergh describes a "democratic deficit" in AI governance when governments or corporations move directly from abstract claims about the common good to technical implementation without sufficient public deliberation [Coe24]. His argument does not deny the importance of expertise. It distinguishes expertise from political legitimacy.

> *Technical expertise can establish what a system can do and help estimate what might happen. It cannot, by itself, determine what society ought to value or what risks others are morally obligated to accept.*

This is particularly relevant to frontier AI because of power concentration. The resources required to train leading models—compute, data infrastructure, energy, specialized chips, capital, and elite technical talent—are concentrated in a relatively small number of corporations. When the same institutions both create the technology and define the acceptable terms of its use, technical concentration can become normative concentration.

Research on U.S. AI policymaking further suggests the need to distinguish healthy industry participation from regulatory capture. Wei et al. explicitly warn that industry involvement is often necessary and beneficial, but document pathways through which concentrated expertise and resources can shape agendas, control information, influence standards, and narrow the range of policy alternatives [Wei24].

The lesson is not that corporate involvement is illegitimate. It is that a governance system relying heavily on corporate expertise needs countervailing sources of expertise and participation if public authority is to remain independent.

8. Intelligence Is Not Wisdom

The terminology of "Super Intelligence" raises a philosophical issue that extends beyond regulation. The following three sections discuss the issue from a Christian perspective.

Modern AI research commonly operationalizes intelligence through performance: prediction, reasoning, optimization, problem solving, planning, tool use, and successful completion of cognitive tasks. Those are genuine and increasingly powerful capabilities. But they do not exhaust the human concept of intelligence.

The Vatican's 2025 note Antiqua et Nova argues that human intelligence is inseparable from embodiment, relationships, moral responsibility, freedom, and the person's orientation toward truth and the good. It cautions against treating computational performance as equivalent to the fullness of human intelligence.

Recent theological scholarship develops this point further. Scherz and Green contrast functional definitions of intelligence as successful rational action with a tradition in which intelligence includes understanding reality, exercising judgment, choosing ends, and acting responsibly in relation to other persons [Sch25].

This difference matters ethically because terminology shapes expectations.

If "intelligence" becomes synonymous with measurable cognitive output, then systems that outperform humans on enough tasks may appear not merely more capable but superior in a more comprehensive sense.

Christian anthropology rejects that inference: Human dignity does not arise from winning a cognitive competition. It does not disappear when a person is less productive, less knowledgeable, disabled, elderly, unconscious, or dependent. It is intrinsic to the person.

> *Machine superiority at a task does not imply moral superiority, greater dignity, or legitimate authority over human ends.*

This observation has practical governance implications.

A system may be better than a physician at detecting a pattern without being capable of bearing moral responsibility for the patient.

It may optimize a logistics network without understanding justice. It may generate persuasive arguments without possessing a commitment to truth. It may recommend a military action without bearing responsibility for death. It may predict human behavior without understanding the dignity of the persons whose behavior it predicts.

The distinction between intelligence and wisdom therefore becomes central. Intelligence, understood as **capability**, expands the set of actions available to society. On the other hand, **wisdom** concerns which actions deserve to be chosen.

The more powerful the capability, the greater the consequences of confusing the two.

8.1 A Biblical Distinction Between Knowledge and Wisdom

The biblical tradition consistently treats wisdom as more than accumulated knowledge, predictive power, or technical mastery. Proverbs 4:7 gives wisdom priority: "Wisdom is supreme. Get wisdom." James 3:17 describes wisdom "from above" by its moral character—peaceful, gentle, reasonable, merciful, and fruitful—while 1 Thessalonians 5:21 calls believers to "Test all things, and hold firmly that which is good." In this framework, the question is not simply whether an intelligent system can produce more powerful answers, but whether the human choices surrounding that system are ordered toward truth, peace, mercy, justice, and the good.

> *Biblically, greater knowledge does not remove the need for wisdom; it increases the responsibility to discern what is good.*

Primary theological source: Antiqua et Nova: Note on the Relationship Between Artificial Intelligence and Human Intelligence

9. Magnifica Humanitas: Human Primacy and the Governance of AI

Pope Leo XIV's encyclical Magnifica Humanitas, signed on 15 May 2026 and promulgated on 25 May 2026, is a central ecclesial text for this paper because it addresses artificial intelligence not as a peripheral ethical issue but as a major social transformation comparable, in its historical significance, to the industrial changes that occasioned Rerum Novarum. Its opening image presents a choice between two forms of human construction: a new Tower of Babel, ordered toward domination and self-exaltation, or a shared work of rebuilding, represented by Nehemiah and the reconstruction of Jerusalem. The question is therefore not whether Christians should reject technological construction, but what kind of social order technological construction is producing (Magnifica Humanitas, nos. 1, 90-94).

The encyclical is not technophobic. It acknowledges technological development as a profoundly human reality and recognizes its capacity to improve living conditions, alleviate suffering, and support integral human development. Its warning is instead directed at the technocratic paradigm: the assumption that efficiency, control, profit, or expanding capability can become sufficient measures of progress. Its concise principle is especially relevant to the rhetoric of superintelligence: "More power does not necessarily imply something better" (no. 93).

9.1 The Primacy of the Human Person

Magnifica Humanitas states that the purpose of moral and social discernment is to safeguard the primacy of the human person, so that human intelligence, conscience, and freedom continue to guide technical innovation and determine its responsible use and limits (no. 97). This is stronger than a general appeal to "human-centered AI." It establishes an ordering principle: technology is an instrument within human moral agency, not an authority that defines its own ends.

Applied to the White House accord, this principle does not exclude corporate self-governance. Developers necessarily exercise first-line responsibility because they possess technical knowledge and operational control. But the primacy of the human person implies that responsibility cannot terminate inside the corporation. Decisions that affect rights, freedom, work, reputation, public goods, or the social distribution of risk require forms of accountability that remain answerable to persons and communities beyond the firm.

9.2 Responsibility, Transparency, and Public Authority

The encyclical's most direct intervention in the self-regulation debate appears in its discussion of responsibility, transparency, and AI governance. It observes that AI decisions affecting employment, credit, public services, and reputation are never purely technical because they touch rights, opportunities, status, and freedom (no. 102). It then argues that ethics in the abstract is insufficient: robust legal frameworks, independent oversight, informed users, and a political system that does not abdicate responsibility are required (no. 106).

This point creates a productive tension with the White House accord. The accord's internal controls, assurance teams, external auditors, and board committees can be understood as serious mechanisms of corporate responsibility. Magnifica Humanitas, however, asks whether these mechanisms are joined to a public architecture capable of establishing enforceable duties, reviewing failures, protecting affected persons, and acting when private incentives diverge from the common good. The issue is therefore not self-regulation versus regulation as mutually exclusive alternatives, but whether self-regulation is nested within institutions of public responsibility rather than substituted for them.

> *Corporate responsibility can be indispensable without being sufficient. Magnifica Humanitas asks whether technological power remains answerable to the human person, the common good, and institutions capable of protecting both.*

9.3 Concentrated Power, Participation, and Subsidiarity

Magnifica Humanitas also widens the governance question beyond model safety. It warns that digital power is increasingly concentrated in actors that control platforms, infrastructure, data, and computing capacity (nos. 95-96). AI can amplify the power of those already possessing economic resources, expertise, and access to data, allowing small groups to influence information, consumption, democratic processes, and economic dynamics (no. 108).

Its response is explicitly participatory. Where AI touches public goods and fundamental rights, oversight should be grounded in participation and subsidiarity; communities and intermediary organizations should not be passive recipients of decisions made elsewhere but should contribute to discernment and oversight (nos. 108-109). The encyclical also treats data governance as a common-good question rather than merely a matter of private ownership.

This reveals an area that the White House accord does not substantially address. The accord focuses on internal controls, safety assurance, auditing, and board governance. Magnifica

Humanitas asks additional questions: who participates in setting the standards, who can contest them, how concentrated technological power is checked, how data is governed, and whether communities affected by AI retain meaningful agency. These are not substitutes for technical safety; they are dimensions of social legitimacy.

9.4 The "More Human" Test: Babel or Jerusalem?

The encyclical ultimately proposes a criterion that aligns closely with this paper's concept of human flourishing. Drawing on Saint John Paul II, it asks whether technology makes human life on earth "more human" and more worthy of the person (no. 129). The alternative is not enthusiasm versus fear but two paths of development: progress that serves persons and peoples, or progress that subjects them to a mentality of power.

The contrast between Babel and Jerusalem gives the governance debate a theological grammar. Babel represents technological coordination directed toward self-magnification, concentration, and domination. Jerusalem, rebuilt under Nehemiah, represents patient reconstruction through distributed responsibility, shared participation, and service to a common life. Applied to AI, the question is not only how powerful systems become, but what form of social relationship their development makes possible.

This suggests a stronger evaluative standard for the White House accord: a governance arrangement should not be assessed only by whether it prevents catastrophic failure. It should also be assessed by whether it keeps human agency primary, subjects concentrated power to accountable oversight, enables meaningful participation, protects the vulnerable, preserves truth and freedom, and directs technological progress toward integral human development.

Principle in Magnifica Humanitas	**Relevant teaching**	**Question for the White House accord**
Primacy of the human person (no. 97)	Human intelligence, conscience, and freedom must guide technical innovation and determine its use and limits.	Do company controls and audits preserve meaningful human authority, including authority outside the developer, over deployment and limits?
Public responsibility and oversight (nos. 102, 106)	AI affecting rights and opportunities is not purely technical; legal frameworks, independent oversight, informed users, and political responsibility are required.	What enforceable public consequence follows when voluntary controls fail, an audit is adverse, or remediation is refused?
Participation and subsidiarity (nos. 108-109)	Communities and intermediary institutions should participate in discernment and oversight where AI affects public goods and rights.	Who beyond companies, auditors, and boards participates in setting acceptable risk, standards, and remedies?
Concentration of digital power (nos. 95-96, 108-109)	Control of data, infrastructure, compute, and platforms can create economic, epistemic, and political asymmetries.	Does the governance framework address concentration and data power, or only model-level safety?

Principle in Magnifica Humanitas	Relevant teaching	Question for the White House accord
Integral human development (no. 129)	The ultimate test is whether technology makes human life more human and more worthy of the person.	Are success metrics limited to capability, safety, and competitiveness, or do they include work, freedom, truth, relationships, justice, and the common good?

Primary theological source: Leo XIV, *Magnifica Humanitas: On Safeguarding the Human Person in the Time of Artificial Intelligence* (2026). Vatican text

10. Human Flourishing as a Governance Criterion

Most AI safety frameworks begin with risks: misuse, malfunction, security, bias, deception, loss of control, or catastrophic harm.

That is necessary but incomplete.

A system can avoid catastrophe while still altering human life in ways that deserve moral examination.

The concept of human flourishing broadens the analysis from "How do we prevent failure?" to "What forms of life does this technology enable or weaken?"

Magnifica Humanitas gives this criterion an explicitly social and theological form. Its question is not whether AI makes human beings "more than human," but whether technological development makes human life more human: more capable of truth, freedom, meaningful work, relationship, participation, justice, care, and responsibility. Human flourishing therefore becomes not an optional ethical supplement to safety, but a test of the ends that safety and innovation are meant to serve (Magnifica Humanitas, nos. 129-131).

Within Catholic Social Teaching and related Christian ethics, several principles are especially relevant.

10.1 A Biblical Vision of Human Flourishing

In Scripture, human flourishing is richer than comfort, productivity, wealth, or technological success. It concerns life rightly ordered toward God, neighbor, justice, peace, fruitfulness, responsible stewardship, and the good of the community. Jesus speaks of his mission in terms of abundant life (John 10:10). Psalm 1:1–3 portrays the flourishing person as a tree planted by streams of water that bears fruit in season. Jeremiah 29:7 instructs God's people to "Seek the peace of the city," linking their own welfare to the welfare of the community. Micah 6:8 summarizes the moral posture required of God's people: to act justly, love mercy, and walk humbly with God. Taken together, these passages make human flourishing a legitimate Christian imperative for evaluating technology: progress should be judged not only by what it enables, but by the kinds of persons, relationships, communities, and institutions it helps to form.

Biblical theme	Selected Scripture	Implication for AI governance
Human dignity	Genesis 1:27; Psalm 8:4–6	Human worth is intrinsic and cannot be ranked by computational performance. AI should remain ordered to the dignity of persons.
Abundant life and shalom	John 10:10; Psalm 1:1–3; Jeremiah 29:7	Assess whether AI contributes to genuinely fruitful lives and the peace/welfare of communities, not merely output or profit.
Wisdom and discernment	Proverbs 4:7; James 3:17; 1 Thessalonians 5:21	Capability must be tested by moral fruits. Technical intelligence should remain subordinate to wise human judgment.
Love of neighbour	Mark 12:31; Philippians 2:4	Governance should consider affected people and externalities, not only the interests of developers, users, or investors.
Justice and mercy	Micah 6:8; Proverbs 31:8–9	Special attention is owed to vulnerable people, unequal burdens, exclusion, and those with limited power to contest automated decisions.
Stewardship and accountability	Genesis 2:15; Luke 12:48	Power is entrusted rather than absolute. Greater capability brings greater responsibility, care, and answerability.
Freedom and non-domination	1 Corinthians 6:12	Human beings should not become mastered by the tools they create; AI should augment rather than erode meaningful agency.
Humility and limits	Mark 8:36; James 4:13–16; Proverbs 16:18	Technological gain is not the highest good. Governance should resist hubris and acknowledge uncertainty, limits, and unintended consequences.
Truth and integrity	Ephesians 4:25; Philippians 4:8	AI systems and their institutions should be oriented toward truthfulness, integrity, and resistance to deception or manipulation.

Scripture quotations and short excerpts in this section use the World English Bible (WEB), a public-domain translation. The biblical references are offered as theological guidance for moral discernment, not as substitutes for technical risk analysis, law, or empirical evidence.

Human dignity

Persons must not be reduced to data points, optimization targets, consumers, labor inputs, or behavioral objects. AI systems should be evaluated partly by whether they preserve the irreducible worth of persons [Lar25, Sch25]. Scripture roots this dignity in the imago Dei: "God created man in his own image" (Genesis 1:27). This means human worth precedes and exceeds measures of intelligence, productivity, or technical competence.

Agency

Human flourishing requires the development and exercise of judgment. If AI systems progressively make decisions on behalf of individuals and institutions, the risk is not only technical error but the erosion of the human capacity to deliberate and take responsibility [Sch25]. Paul's warning that he would not be "brought under the power of anything" (1 Corinthians 6:12) provides a useful principle of non-domination: tools should serve responsible human agency rather than quietly mastering it.

Subsidiarity

Decisions should be made at the most local competent level, with higher levels supporting rather than unnecessarily displacing lower-level agency. Highly centralized AI infrastructure can place extraordinary informational and decision-making power in a small number of institutions. Subsidiarity therefore provides a lens for examining not only model behavior but institutional concentration [Sch25]. The New Testament image of one body with many members (1 Corinthians 12:12–26) offers a complementary biblical picture in which unity does not erase distinct roles, local responsibility, or mutual dependence.

Solidarity

The benefits and burdens of AI are unlikely to be distributed evenly. Workers whose roles are automated, communities hosting energy-intensive infrastructure, people subject to automated decisions, and countries lacking access to frontier systems may experience the same technology differently. A flourishing framework therefore asks who benefits, who bears risk, and who participates in the decisions. Jesus' command, "You shall love your neighbor as yourself" (Mark 12:31), and Philippians 2:4's concern for the interests of others direct attention to people who bear costs without sharing proportionately in the benefits.

Stewardship

Technological capability is understood as something to be used responsibly rather than as an end in itself. "Can it be built?" and "Can it be deployed?" remain different questions from "What human purpose does it serve?" Genesis 2:15 depicts humanity as entrusted to cultivate and keep creation, while Luke 12:48 states the accountability principle that to whom much is given, much is required. Greater technological power therefore entails greater moral responsibility.

These principles do not supply a ready-made statute. Their contribution is normative. They prevent governance from collapsing into a single objective such as maximizing innovation, minimizing immediate harm, or winning geopolitical competition.

Capability asks	What can we do?
Safety asks	What might go wrong?

Law asks	What is permitted?
Ethics asks	What ought we to do?
Wisdom asks	What kind of persons and society are we becoming?

A governance system focused only on the first three questions may be technically sophisticated and still morally incomplete.

11. Governance Alternatives and Their Trade-offs

The policy debate is often framed as a binary choice between "innovation" and "regulation," or between "self-regulation" and "government control." The study of the literature suggests a wider institutional design space. The following models can be analyzed without assuming that one arrangement is universally superior.

Model	Core design	Potential advantages	Potential vulnerabilities
Voluntary corporate governance	Companies establish safety frameworks, employ internal controls, commission external audits, and rely on board oversight.	Speed, technical expertise, adaptability, lower bureaucratic burden, strong ownership of safety by developers.	Inconsistent implementation, weak verification, limited transparency, commercial conflicts, race dynamics, and absence of enforceable consequences.
Co-regulation	Government establishes broad legal duties while accredited private or multi-stakeholder bodies develop technical standards and conduct assessments.	Combines public authority with specialist expertise; adaptable standards can evolve faster than statutes.	Complexity, capture of standards bodies, uneven auditor quality, and uncertainty about responsibility when private standards fail.
Principles-based public regulation	Law establishes high-level duties and gives a specialist regulator information-gathering and enforcement powers; companies retain flexibility in how they satisfy those duties.	Flexibility with enforceability; places technical innovation on developers while retaining public oversight.	Demands highly capable regulators; broad principles may create uncertainty or discretionary enforcement.

Model	Core design	Potential advantages	Potential vulnerabilities
Risk-tiered statutory regulation	Legal obligations increase with model capability, deployment context, compute scale, or demonstrated risk.	Avoids imposing frontier-level burdens on low-risk systems; targets systems most capable of large-scale harm.	Thresholds can be gamed or become obsolete; capability measurement remains difficult; compliance costs may favor incumbents.
Polycentric governance	Corporations manage internal safety; auditors verify; governments establish legal baselines; universities, civil society, standards bodies, and international institutions govern different dimensions.	No single institution bears the entire epistemic or moral burden; multiple centers can correct one another's failures.	Fragmented responsibility, conflicting standards, slow coordination, and opportunities to shift accountability among institutions.

Schuett et al. describe principles-based regulation as potentially appropriate during periods of technological uncertainty, with more specific rules developing as evidence and safety practice mature [Sch24].

> *Which decisions require technical discretion, which require independent verification, which require public authority, and which require democratic or moral deliberation?*

Different answers may be appropriate for different levels of risk.

12. A Framework for Evaluating the White House Accord

Rather than assigning a political verdict to the September 29 initiative, it can be evaluated through a set of institutional questions.

Test	Question
1. Epistemic adequacy	Does the system reliably discover known and unknown risks? Are evaluations capable of measuring dangerous capabilities before deployment? Are thresholds measurable?
2. Verifiability	Can parties outside the company verify that commitments have actually been implemented? Are claims supported by accessible evidence?
3. Auditor independence	Are evaluators financially, organizationally, and professionally independent? Do they have sufficient access? Who accredits them?

Test	Question
4. Consequence	What happens if controls fail, an audit is adverse, or management declines remediation? Is there a mechanism beyond reputational pressure?
5. Transparency	Which findings are reported to regulators, users, affected communities, and the public? What information can legitimately remain confidential for security or proprietary reasons?
6. Contestability	Can affected people challenge decisions made with or by advanced AI systems? Are there rights of appeal in high-stakes domains?
7. Democratic legitimacy	Who defines acceptable societal risk? Are decisions confined to developers and technical experts, or are public institutions and affected communities represented?
8. Distribution	Who receives the benefits of frontier AI and who carries its costs, including labor disruption, data-center burdens, surveillance, and security risks?
9. Human agency	Does the technology augment human judgment or gradually substitute for it in ways that diminish responsibility and practical wisdom?
10. Adaptability	Can the governance system tighten, loosen, or change requirements as evidence develops without becoming either rigid or arbitrary?
11. Christian moral discernment	Does the governance arrangement serve life, human dignity, truth, justice, mercy, neighbour-love, stewardship, responsible agency, peace, and humility? Does it satisfy the Magnifica Humanitas test of making human life more human, while preserving participation, subsidiarity, and accountable human authority? Does it test technological gain against the possibility of moral or human loss (Mark 8:36)?

This framework makes it possible to analyze the accord empirically as implementation develops. It also separates questions that are sometimes conflated. A company can score highly on internal technical safety while questions of legitimacy or distribution remain unresolved.

For a Christian evaluation, the eleventh test is not an alternative to the preceding institutional tests but a moral horizon for them. Scripture asks what the exercise of power is ultimately for, whom it serves, what kinds of fruit it produces, and whether apparent gain comes at the cost of human dignity, truth, justice, responsible freedom, or love of neighbour. Magnifica Humanitas makes that horizon institutionally concrete: the primacy of the human person requires accountable human authority, effective oversight, participation, subsidiarity, attention to concentrated power, and a conception of progress measured by integral human development rather than capability alone.

13. Conclusion: The Governance Problem Behind the Intelligence Problem

The September 29 actions crystallize a dilemma that will likely outlast the particular terminology of "Super Intelligence."

The United States is simultaneously pursuing technological leadership, rapid innovation, frontier-system safety, and public trust. These goals are not inherently incompatible. They can, however, come into tension when safety measures slow deployment, when transparency conflicts with security, or when public regulation appears to threaten competitive advantage.

The White House accord recognizes that frontier AI requires structured governance. Its four layers, namely internal controls, internal assurance, external evaluation, and board oversight, are serious institutional mechanisms, not merely slogans. The open question is whether they can carry the full weight placed upon them when participation is voluntary, standards remain partly unspecified, public verification is limited, and the companies involved operate under intense competitive and geopolitical pressure.

The existing scholarship offers two conclusions that can coexist:

a. Frontier AI governance cannot work without industry. Developers possess essential technical knowledge and must maintain sophisticated internal safety systems.
b. Frontier AI governance cannot be reduced to industry. Decisions about acceptable public risk, rights, distribution, human agency, and the purposes of transformative technologies reach beyond corporate expertise.

The deepest question is therefore not simply whether an artificial system becomes "superintelligent."

It is whether human institutions remain capable of governing powerful technologies without surrendering the distinction between capability and wisdom, expertise and authority, innovation and the common good.

If future systems exceed human beings across an expanding range of cognitive tasks, the moral status of human persons does not diminish. On the contrary, the need to articulate what is distinctively worth preserving in human agency becomes more urgent.

Magnifica Humanitas gives the same concern a social-doctrinal form. Its warning is not that technological power is evil, but that power can grow without a corresponding growth in humanity. The appropriate question is therefore not only whether AI can be made more capable or even more secure, but whether its institutions of development and governance make human life more human — more free, truthful, just, participatory, relational, and oriented toward the common good (Magnifica Humanitas, nos. 93-96, 129).

Scripture supplies a demanding limit to any technological triumphalism. Jesus asks, "For what does it profit a man, to gain the whole world, and forfeit his life?" (Mark 8:36). Applied analogically to technological civilization, the warning is clear: greater capability, wealth, power,

or competitive advantage cannot by themselves define the human good. Proverbs 4:7 places the priority differently: "Wisdom is supreme. Get wisdom." Intelligence and technological power are instruments; wisdom must govern the ends toward which they are directed.

> *The more powerful our machines become, the more carefully human beings must distinguish what intelligence can accomplish from what wisdom has reason to choose.*

That is a question no model, corporation, administration, legislature, church, or profession can answer alone.

References

Primary source Executive Order 14434. "Inaugurating the Era of Super Intelligence." The White House, 29 September 2026. Link

Primary source "White House Accord on Super Intelligence: Joint Commitment on Frontier Responsibilities." 29 September 2026. Link

[Wan25] Wang, Jennifer, Kayla Huang, Kevin Klyman, and Rishi Bommasani. "Do AI Companies Make Good on Voluntary Commitments to the White House?" 2025. https://arxiv.org/abs/2508.08345 Link

[Ste25] Stelling, Lily, et al. "Evaluating AI Companies' Frontier Safety Frameworks: Methodology and Results." 2025. https://arxiv.org/abs/2512.01166 Link

[Fer21] Ferretti, T. "An Institutionalist Approach to AI Ethics: Justifying the Priority of Government Regulation over Self-Regulation." Moral Philosophy and Politics (2021): 239–265. Link

[And23] Anderljung, Markus, et al. "Frontier AI Regulation: Managing Emerging Risks to Public Safety." 2023. https://arxiv.org/abs/2307.03718 Link

[Sch24] Schuett, Jonas, Markus Anderljung, A. Carlier, Leonie Koessler, and Ben Garfinkel. "From Principles to Rules: A Regulatory Approach for Frontier AI." In *The Oxford Handbook of the Foundations and Regulation of Generative AI,* Oxford University Press, 2025. Link

[Wei24] Wei, Kevin, Carson Ezell, Nick Gabrieli, and Chinmay Deshpande. "How Do AI Companies 'Fine-Tune' Policy? Examining Regulatory Capture in AI Governance." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (2024): 1539–1555. Link

[Coe24] Coeckelbergh, Mark. "Artificial Intelligence, the Common Good, and the Democratic Deficit in AI Governance." AI and Ethics (2024): 1491–1497. Link

[Bru26] Brundage, M., et al. "Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies." 2026. https://arxiv.org/abs/2601.11699 Link

[Cos22] Costanza-Chock, Sasha, et al. "Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem." ACM FAccT (2022). Link

[Sch25] Scherz, P., and B. Green. "Reclaiming Human Agency in the Age of Artificial Intelligence." Journal of Moral Theology (2025). Link

[Lar25] Laracy, Joseph R., et al. "Human Dignity and the Ethics of Artificial Intelligence: A Framework for Responsible Design and Use from the Perspective of Catholic Social Teaching." IEEE ETHICS (2025): 1–10. Link

Theological source Dicastery for the Doctrine of the Faith and Dicastery for Culture and Education. Antiqua et Nova: Note on the Relationship Between Artificial Intelligence and Human Intelligence. 28 January 2025. Link

Theological source Leo XIV. *Magnifica Humanitas: On Safeguarding the Human Person in the Time of Artificial Intelligence.* Signed 15 May 2026; promulgated 25 May 2026. Link

Scriptural sources World English Bible (WEB), public domain. Principal passages discussed: Genesis 1:27; 2:15; Psalm 1:1–3; 8:4–6; Proverbs 4:7; 16:18; 31:8–9; Jeremiah 29:7; Micah 6:8; Mark 8:36; 12:31; John 10:10; 1 Corinthians 6:12; 12:12–26; Philippians 2:4; 4:8; 1 Thessalonians 5:21; James 3:17; 4:13–16; Luke 12:48; Ephesians 4:25.