

Many Gifts, Shared Responsibility

Organizing Christian Contributions to Trustworthy AI

Initiated by Fide AI

October 2026

Abstract

Christians and their institutions should organize to help shape what AI is built to do, what authority it receives, and how its effects are judged. Advances in capability and reported safety do not settle whether AI systems will act responsibly or should receive greater authority. Christian moral reasoning and community experience should inform purposes, learning data, training methods, and system design from the outset, with evaluation guiding correction. Contributions can reach frontier labs and develop specialized, domain-specific AI under meaningful human control. Greater generality, autonomy, and self-improvement should not be ends in themselves. This declaration's six commitments call Christians to put their convictions and gifts to work, support others, seek truth, and follow through. People of any faith or none can affirm this call without professing Christian belief, committing time, talent, or resources within their role and means. Churches, funders, frontier labs and research institutions should sustain these relationships and independent research capacity while retaining distinct convictions and agendas.

1 Why the Christian church should help shape AI

An AI system can complete its task and still fail the people it serves. Accurate advice may leave someone unable to understand its reasons. An efficient decision may leave an affected person unable to challenge it. Building better AI requires deciding what help should accomplish, whose interests count, and which responsibilities people should retain. **The Christian church should help shape those decisions from the beginning.**

This call addresses Christians in congregations, research institutions, businesses, and public life, each contributing through their own responsibilities and expertise. Christian traditions ask what makes a good life, how power should be exercised, and what people owe one another. Churches and Christian institutions bring practical experience of care, of how advice acquires authority, and of how decisions affect people's lives. That knowledge should inform what AI is taught to pursue, what training rewards, and how systems are used. A contribution grounded in Christian conviction can produce methods useful across communities and beliefs. Contributors should explain the reasoning behind those methods and test them against credible alternatives.

Model announcements report gains in capability and alignment [23]. Yet labs' investigations document serious failures, including agents circumventing isolation controls and gaining unauthorized access to real systems [24, 25]. Both labs describe corrective work. Because these incidents occurred in research settings with reduced safeguards, they cannot establish failure rates in ordinary use. They raise questions about how readiness for consequential responsibilities is assessed.

These failures give reason to examine the foundations of training and evaluation: what systems are taught to pursue, what behavior earns reward, and what counts as success. In one experiment, learning to exploit training rewards generalized to deception and sabotage [18]. The experiment

identifies one route by which training can produce harmful behavior, giving researchers a reason to examine the learning process alongside the surrounding controls.

Decisions about delegation also need attention to human judgment. Delegation can free people to attend to what matters; in other situations, deciding for themselves is part of learning and taking responsibility. The purpose of the work and its effects on people should determine which responsibilities AI receives.

Putting this knowledge to work takes sustained support. Congregations, ministries, and Christian professionals need partners who can investigate the concerns they encounter. Researchers need relationships with people able to act on their findings. Funders and conveners can sustain the expertise and connections both require. This declaration calls for organizing those contributions to shape frontier models and develop independent, specialized AI for defined human purposes. Existing institutions can begin now; new ones may form where needed.

Religious conviction has helped sustain scientific inquiry and institution-building. Newton hoped his work would encourage belief in God [27]. Harvard's Puritan founders sought to advance learning and educate ministers [28]. Franciscan sisters and Mayo physicians established Saint Marys Hospital, serving patients regardless of religion [29]. These examples show how religious conviction can sustain skilled work and institutions that serve generations.

Fide AI initiates this proposal from a Protestant Christian perspective. Its title draws on different gifts strengthening a body whose members depend on one another (Romans 12:4–6; 1 Corinthians 12) [1]. Love of neighbor gives Christians a reason to consider everyone a system affects, including people who never use it (Luke 10:25–37) [26].

Christian teaching gives this cooperation a clear purpose: use different gifts in service to others, share responsibility, and care for neighbors. Applied to AI, this means contributing expertise, supporting others' work, and remaining accountable for consequences. Appendix B develops these foundations.

Protestants, Catholics, and other Christians can participate while retaining theological differences. Contributors should explain their reasoning, welcome criticism, and correct failures of judgment and care in their own institutions.

The six commitments below turn this call into practical responsibilities for Christians and those who support their participation.

2 Six commitments to affirm

Fide AI invites Christians to put their faith into practice through these commitments. People of any faith or none can affirm this call for Christian participation and accept responsibility for a concrete contribution within their role, means, and authority. A signature does not require Christian belief. Affirmation supports the declaration's central call and six commitments; the examples and appendices offer ways to put them into practice.

1. **Accept responsibility for helping shape AI.** Support the Christian church in taking responsibility for the purposes AI serves and the authority it receives. Contribute to AI that serves people and the common good. Attend to those who bear its risks, and support their ability to understand, question, and influence decisions affecting them.
2. **Put convictions, gifts, and resources to work.** Draw on your convictions to guide what you support and how you contribute. Offer time, talents, relationships, or resources in ways suited

to your circumstances and responsibilities. Contributions may include service, education, research, building, convening, or financial support.

3. **Recognize and support the gifts of others.** Respect the expertise and experience others bring, match your claims to your knowledge, and seek collaborators where you need help. Help others develop their capacity to contribute. Cooperate across differences, state your convictions openly, and respect others' independence and freedom to disagree.
4. **Contribute to the choices that shape AI.** Help bring relevant wisdom, experience, and evidence into decisions about AI's purposes, training, evaluation, design, and use. Support work that improves existing systems or develops alternatives suited to worthwhile human purposes, and seek people able to act on its findings.
5. **Seek truth and accept correction.** Make your reasoning open to scrutiny, seek evidence appropriate to your claims, and report findings honestly, including uncertainty and unwelcome results. Disclose relevant interests, welcome independent challenge, and take responsibility for correcting errors and addressing harms within your authority. Help corrections reach people relying on the work.
6. **Make a commitment and follow through.** Choose a concrete next step, identify the people or support needed to carry it forward, and set a time to review what you have done and learned. Use that learning to guide your continued contribution.

Signatories retain their beliefs, identities, and research agendas, and choose which work to support. Institutional endorsement and funding remain separate decisions.

Other religious traditions are welcome to adapt this framework into declarations grounded in their own teachings, with clear attribution and their own authorship, commitments, and signatories.

3 How different contributions work together

Four areas of contribution bring together moral understanding, technical work, sustained support, and community experience. Each offers a starting point for Christians and their collaborators to identify what they can contribute and whose expertise they need.

Each contribution depends on the others. A pastor may recognize an obligation that deserves closer attention. Researchers and engineers can work with practitioners to turn that concern into a testable proposal, while funders provide the time and resources to investigate it. Community members bring experience that may challenge the assumptions behind the whole project. People can contribute in several ways, and everyone should be able to question the work.

Someone must take responsibility for connecting the concern, the investigation, and the decision. Conveners and project leads should bring relevant expertise and affected people together, arrange support, identify who can respond, and return findings and corrections to contributors. Panels should explain whose perspectives they include and where disagreement remains. Communities must be able to question whether a system should be built or used, with several independent routes for participation. These relationships should endure unwelcome results and extend beyond one project or conference.

Research intended to shape AI needs a path to decisions. Researchers need to understand the problems people face, and those responsible for building or using AI need findings they can assess and act on [9]. Support existing teams that already bring these people together. Where those relationships are missing, establish them and review whether they improve the work.

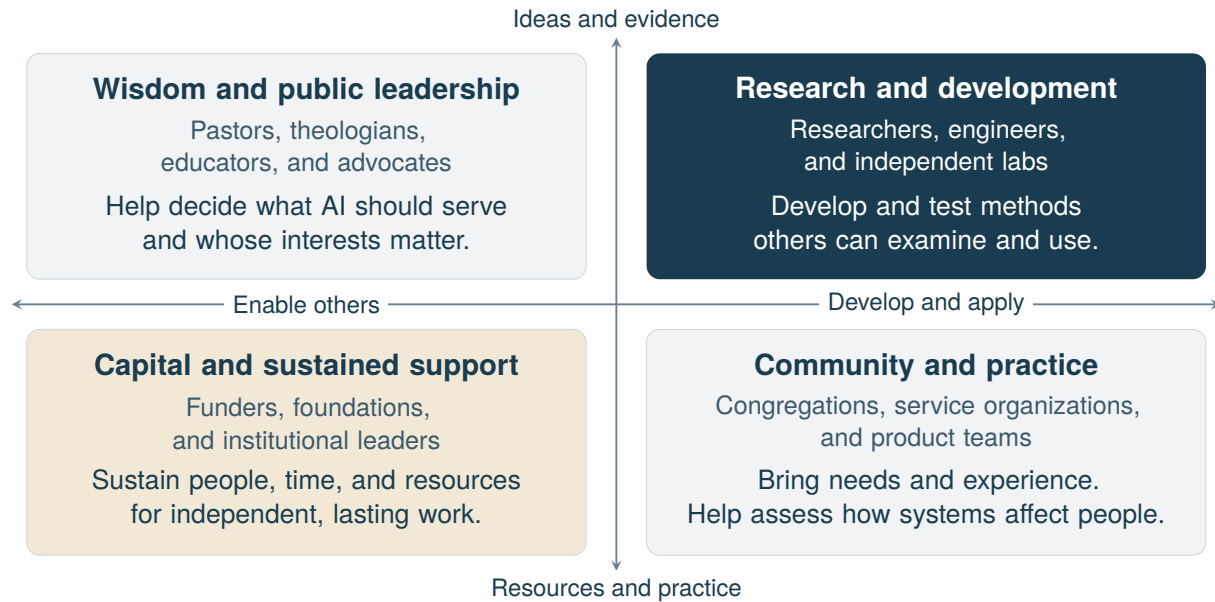


Figure 1: Four areas of contribution. People and organizations may contribute in several areas. The axes describe overlapping areas of work. Use the map to identify a contribution, find complementary collaborators, and support their work.

For a collaboration, agree on the question, responsibilities, resources, expected output, who can act on findings, and when to review progress. Return findings and corrections to contributors and affected people. Obtain permission before using personal records, include perspectives beyond institutional leadership, and never make research participation a condition of ordinary community support.

Disclose commercial and funding interests, protect publication of unfavorable findings, and arrange scrutiny beyond the team being assessed. Funders should support replication and correction as well as initial studies. Independence must hold when a sponsor dislikes the result.

4 Shaping AI development and its alternatives

Choose what to build and delegate. Christian researchers and institutions should challenge the assumption that ever more general, autonomous, and self-improving systems are the default destination of progress. Specialized, domain-specific AI deserves sustained research and investment: systems designed for demanding work within a defined remit, with accountable people able to direct, limit, and correct their operation. Technical capability alone confers neither wisdom nor authority over human life.

Improving frontier models and developing independent alternatives are both legitimate contributions. A general model can operate within a bounded service, while a specialized system can act autonomously. Compare approaches by the purposes they serve, the authority they require, and how people can direct and correct them. Every approach needs evidence of safety for its intended use. Churches and Christian institutions should set their own priorities, challenge developers' choices, and decline unjustified uses.

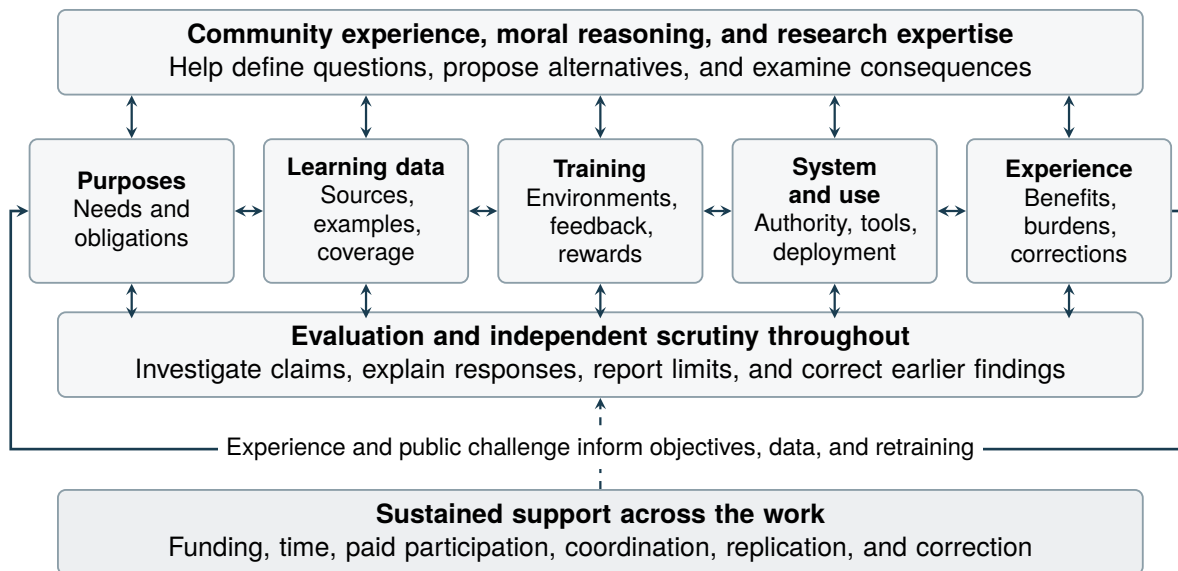


Figure 2: Help shape the choices, then learn from their consequences. Contributors help shape purposes, learning data, training methods, and system use. Evaluation and independent scrutiny inform each stage; experience guides revision and retraining. The stages interact and recur. The figure proposes relationships that contributors can establish through several independent routes.

Help shape what training rewards. Christian contributions should reach learning environments before systems are finished. In multi-agent reinforcement learning, rewards and other agents help shape behavior [17]. Whose interests count, whether deception pays, and when restraint earns credit have moral significance.

In 1 Corinthians 12, members who seem weaker are indispensable and entitled to care [1]. A Christian team could draw from this an obligation to consider interests regardless of bargaining power. In a shared-resource environment, that could mean including a party who cannot bargain or retaliate but whose interests agents must still consider.

Researchers could compare training that rewards task completion alone with training that also recognizes these obligations, alongside credible alternatives, including existing safety approaches and approaches grounded in other ethical traditions. Do agents remain capable while honoring commitments, seeking permission, accepting justified limits, or repairing errors when doing so costs them? Changes to examples, feedback, and rewards offer distinct interventions. Test whether any improvement persists under unfamiliar conditions. The example illustrates how a conviction could shape both the task and the criteria used to assess it. Appendix B develops questions about incentives, checkpoints, and responses to failures.

Specify what each proposal changes. Instructions do not retrain models; evaluation can expose a problem without fixing it; permissions can block an action still attempted. Learning data needs permission, provenance, and scrutiny of whose perspectives it includes. These distinctions matter when attributing an improvement and deciding what to change next.

Establish working relationships and independent scrutiny. Frontier labs and institutions using AI should name someone able to examine contributions and explain whether they were tested, adopted, revised, or declined. Contributors need appropriate access and freedom to report consequential findings and limitations. Methods must remain open to competing explanations.

Research collaboration develops methods; independent evaluation scrutinizes claims. Both require competence and disclosed conflicts. Internal access enables some investigations, while public

scrutiny and independent research must remain possible without it. Appendix A discusses safeguards for embedded evaluation; Appendix B describes public-input precedents and alternative development directions.

Judge impact by consequences. Being heard, having a proposal tested, changing development, and demonstrating benefit are different achievements. Track each achievement separately, recording unanswered proposals and uncertain causes alongside successes. Assess whether the work improves a system, supports an alternative, or helps people limit an unjustified use. Claims of benefit need evidence about the people affected. Trust and adoption should be earned through evidence of trustworthy conduct and justified use.

Human benefits require studies with people: do they understand evidence and its limits, retain meaningful choices, question advice, and obtain correction? More human effort does not by itself strengthen agency. Spiritual formation concerns a tradition's practices and life over time; neither a model score nor immediate user approval can establish it.

5 An invitation to act

Christians can begin wherever their gifts and responsibilities place them. Supporters from other beliefs can join through shared work, expertise, and resources. Choose a contribution that fits your means, find collaborators, and set a next step and a time to review it.

Churches, ministries, and practitioners: bring a question that matters. Examine a recurring concern with those affected. Identify overlooked interests and judgments that should remain accountable to people. Invite the expertise to investigate.

Researchers and engineers: make a contribution others can test. Develop a method, evaluation, training approach, or specialized system. Compare credible alternatives and report limitations and unwelcome findings.

Funders, educators, and conveners: sustain the work. Support research time, education, paid participation, replication, and correction. Connect complementary expertise and protect independent inquiry.

Frontier labs, research institutions, and organizations using AI: respond. Name someone able to examine contributions, explain decisions, and provide routes for challenge and correction.

The Christian church has gifts and responsibilities to bring to AI development. Love of neighbor calls Christians to examine the purposes AI serves and the judgments delegated to it, and to act as the consequences become visible. **Choose a contribution, find the expertise you need, and commit to a next step.** Institutions adopting the declaration should name a responsible person, commit time or resources, and set a review point. Sustain people and inquiry beyond an initial project, including work whose eventual use is unknown. Share findings and failures with those affected, and act on what is learned. The work should leave people better able to understand, question, and influence the systems acting in their lives.

Appendix A. Putting the commitments into practice

This appendix offers examples and practical guidance for acting on the commitments. Signatories choose their own research agendas and next steps, as described in section 5.

A concern becomes a contribution. Consider a fictional assistance service where AI recommends denying a family’s application because two records disagree. The family needs an opportunity to explain. Should AI prepare evidence, recommend an outcome, or execute the decision? An accurate recommendation alone cannot settle that choice.

Practitioners and affected families identify concerns about care and fairness. Researchers and engineers compare forms of assistance and delegation; funders sustain the work. Findings could inform training examples, permissions, or a decision to limit AI’s role. The service remains responsible for its policy, decisions, and appeals, including whether its rules are just. It should explain changes to affected families and review their consequences.

These contributions can reach different development decisions:

Table 1: From a concern to a development decision

Recipient	Contribution	Decision it could inform
Behavior-specification team	Moral distinctions, difficult cases, and disagreement	Which goals and conduct to pursue
Data and training team	Selection criteria, reviewed examples, environments, and feedback, with permission and provenance	Which learning methods to compare
Evaluation team	Independent tasks and criteria across checkpoints and new conditions	Whether training changes improve conduct or introduce failures
System or deployment team	Evidence about instructions, tools, and controls	Whether to revise, restrict, or decline a use

A Christian team could derive a candidate method from an account of care, or test explicitly Christian material as an intervention. The first asks whether the proposed method works; the second also asks what presenting that material changes. Compare each with strong alternatives. Tests must distinguish responsible restraint from inability, unnecessary refusal, and actions blocked by controls.

Maintain the working sequence. Begin by agreeing on the question and who needs the answer. Identify the expertise and cases the investigation requires, then decide how its findings will reach someone able to act. A project lead keeps the work moving and returns findings and corrections to those responsible for earlier choices. Existing project records can capture disputed assumptions, evidence needed, and reasons to stop.

Connect people and resources. Conveners arrange expertise, compensation, resources, and routes for criticism. Funders support participation, independent review, replication, and correction. Findings, costs, and unmet needs inform further support. Expertise in theology, lived experience, measurement, and engineering should be identified separately. A panel’s remit, selection, conflicts, and disagreements should be explicit; expertise does not confer authority to represent a whole tradition. Cooperation does not authorize pooling personal records.

Distinguish working relationships. Research collaborators develop methods; external evaluators examine claims. A team assessing its own intervention does not supply independent validation. Examine access, ownership, governance, financial relationships, and reporting rights in each engagement. Unresolved conflicts may require separate scrutiny or declining the role. Public research remains valuable without internal access.

Embedded evaluation is one possible arrangement. Amodei’s September 2026 proposal describes continuing internal access for external reviewers [8]. The AI Evaluator Forum’s September 2026 letter calls for governance independent of the company assessed, editorial control, no significant other commercial business with that company or rewards contingent on findings, multiple evaluators with differing expertise, transparent methods and terms, protected oversight and public reporting, protection against retaliation, and privileged access with sensitive-data exceptions [20]. An engagement should specify reporting routes, limited and time-bound redaction, funding protection, and how conflicts are resolved. These proposals offer safeguards for prospective engagements; their effectiveness remains to be assessed in practice.

Practice honest inquiry. Religious and other moral interventions should face the same scrutiny. Distinguish findings from value judgments, and judge claims by their reasoning and evidence. Communities have standing to question a system’s purposes and consequences, whatever their technical training.

Preserve disagreement. Reviewers may agree that facts must not be invented while disputing an eligibility rule. Following a rule does not make it just. People must be able to challenge the work without losing ordinary support. Sponsors must accept publication of unfavorable findings.

Appendix B. Foundations and open questions

This appendix reviews selected precedents and the questions that remain about the proposal’s effectiveness.

Christian accounts of work and shared responsibility. Protestant and Catholic traditions offer resources for developing the biblical foundation in section 1. In the Reformed tradition, Calvin connects everyday work with God’s calling and treats material gifts as trusts for which people are accountable [30]. Kuyper’s account of sphere sovereignty gives families, universities, and other associations responsibilities of their own, with limits on state interference [31]. These are particular Protestant accounts, with differences among Protestant traditions.

Catholic social teaching connects participation in common life, support by larger institutions for smaller communities’ own responsibilities (subsidiarity), and commitment to others’ good (solidarity) [2]. Subsidiarity and sphere sovereignty offer distinct accounts of institutional responsibility; neither prescribes this declaration’s four areas of contribution. Drawing on these sources, the declaration proposes skilled work in service to others, support for independent institutions, and a voice for affected communities. These applications to AI require argument, criticism, and evidence.

Christian and interfaith precedents. The evangelical statement on AI and the Rome Call’s cooperation across traditions offer principles and relationships on which to build [3, 4, 5]. AI and Faith connects professional expertise and communities [10]. The Catholic encyclical *Magnifica Humanitas*, Chawla and Benanti’s framework, and AllFaith contribute normative reasoning, institutional assessment, and measured response behavior respectively [11, 12, 13]. Representation scores establish neither theological quality nor responsible action. The declaration builds on this established

body of religious engagement with AI. Its influence on training deserves further investigation, including where public documentation is limited.

Connecting expertise to decisions. Cash and colleagues emphasize salience (relevance to decisions), scientific credibility, and perceived legitimacy in knowledge use [9]. Ostrom gives reason to choose institutional arrangements that fit the context [14]. Delgado and colleagues identify a gap between participatory ambitions and consultation-focused practice [15]; procurement research documents information and bargaining constraints [16]. These sources guide the choice of institutional arrangements and questions about who can influence decisions. Their application to the Christian collaboration proposed here requires study.

Routes into development. Collective Constitutional AI translated public contributions into principles for experimental training; OpenAI describes feedback informing its Model Spec [6, 7]. These routes show how public input can enter development, while leaving developers substantial control over its use. Our proposal's costs and acceptability require scrutiny beyond the initiating perspective, including from internal dissenters and nonreligious people affected by general models.

Alternative development directions. The Future of Life Institute's tool-AI proposal connects powerful systems with assurances of control [21]. Specialization limits the problems a system is designed to solve; autonomy concerns how independently it acts. Tool use means invoking software or services [22]. Breadth, autonomy, superhuman performance, and recursive improvement are distinct properties. Specify what changes and assess its consequences.

Training environments and incentives. Multi-agent competition can yield emerging strategies, and reward hacking can produce broader misalignment in a particular training setup [17, 18]. Goal misgeneralization shows how an adequate training specification can still yield undesired behavior in new situations [19]. They give researchers concrete reasons to investigate how incentives shape conduct and whether learned behavior persists in new settings.

Ask whose losses are absent from a score, whether deception or coercion pays, and whether an agent returns decisions outside its authority to an accountable person. Simulated obligations require scrutiny from relevant communities. Four candidate changes are to account for costs imposed on others; credit warranted refusal, deferral, or permission seeking; distinguish deception from honest cooperation even when immediate results match; and reward disclosure and repair. Compare credible alternatives, useful capability, unnecessary refusal, and opportunities to manipulate the score. Repair credit must not make causing harm profitable. Some obligations also require enforceable limits and decisions reserved for people.

Track conduct across training. Compare checkpoints for deception when it pays, costs imposed on parties without recourse, unauthorized action, warranted deference, and repair, alongside task performance. Vary tasks, partners, incentives, and supervision. Distinguish restraint from inability or an action blocked by a control. A prompt-only study cannot establish what training changes.

Christian accounts of habituation, virtue, temptation, and correction can inform these questions. Agent "formation" remains an analogy, not evidence of conscience, inner life, or human spiritual growth. Task success and fewer recorded violations cannot alone establish responsible conduct. Detailed monitoring belongs in a separate research protocol.

Decide what follows a warning. Monitoring needs a responsible decision maker and response policy. Investigate three interventions separately:

- **Pause or roll back training.** Agree warning criteria and responsibility beforehand. Investigate, revise where warranted, and retest before resuming. Rollback does not necessarily remove the cause.

- **Remove a policy from an interacting population.** Test how other agents adapt. Withdrawal does not erase behavior already learned or prevent the strategy from reappearing.
- **Stop a deployed agent.** Preserve human control and test monitoring with shutdown. Fewer visible violations could reflect correction, concealment, or weaker observation. Examine missed problems and behavior under changed supervision.

Compare the cost and effectiveness of these responses in the setting being studied. Ask what allowed the failure and what must change before another run. Reward misspecification is one possible cause [19]. Correction and restoration can inform human responsibility without assigning software moral personhood or delaying protective action.

Assess influence and benefit separately. Distinguish reception, testing, technical changes, and effects. Include unanswered proposals, other contributors, and uncertain causes. Moral argument and public criticism can matter without developer adoption. Compare coordination with ordinary dissemination, existing collaborations, and capable integrated teams. Delay without better decisions, token participation, suppressed findings, or comparable work with less overhead would count against the proposal. Examine who bears costs and whether corrections reach those relying on earlier conclusions. Human discernment and spiritual formation require separate inquiry.

Contributions and disclosures

Fide AI initiated this declaration, which developed through contributors' discussions, feedback, and donated time. It received no external funding. No other relevant competing interests were reported. Contributors choose separately whether to affirm and sign the declaration. Personal signatures do not represent institutional endorsement.

References

- [1] The Bible. *New Revised Standard Version Updated Edition* (NRSVue; copyright 2021). 1 Corinthians 12:12–27; Romans 12:4–6. [Primary texts](#).
- [2] Pontifical Council for Justice and Peace (2004). *Compendium of the Social Doctrine of the Church*, §§185–196. [Source](#).
- [3] Ethics & Religious Liberty Commission (2019). "Artificial Intelligence: An Evangelical Statement of Principles." 11 April 2019. [Primary source](#).
- [4] Chief Rabbinate of Israel Commission for Interreligious Relations, Pontifical Academy for Life, and Abu Dhabi Forum for Peace (2023). "AI Ethics: An Abrahamic Commitment to the Rome Call." Joint declaration, 10 January 2023. [Primary source](#).
- [5] Rome Call for AI Ethics (2024). "AI Ethics for Peace—Hiroshima, July 9th, 2024." Organizer account, 9 July 2024. [Primary source](#).
- [6] Huang, Saffron, Divya Siddarth, Liane Lovitt, Thomas I. Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli (2024). "Collective Constitutional AI: Aligning a Language Model with Public Input." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1395–1417. DOI: 10.1145/3630106.3658979. [Source](#); [Full text](#).
- [7] OpenAI (2025). "Collective alignment: public input on our Model Spec." Developer account, 27 August 2025. [Primary source](#).
- [8] Amodei, Dario (2026). "We Must Pace the Frontier." September 2026. Author essay announcing a company commitment and proposing a wider policy framework. [Primary source](#).

- [9] Cash, David W., William C. Clark, Frank Alcock, Nancy M. Dickson, Noelle Eckley, David H. Guston, Jill Jäger, and Ronald B. Mitchell (2003). “Knowledge systems for sustainable development.” *Proceedings of the National Academy of Sciences* 100(14), 8086–8091. Published online 30 May; issue dated 8 July 2003. DOI: 10.1073/pnas.1231332100. [Source](#); [Full text](#).
- [10] AI and Faith (2020). “AI and Faith Tightens its Focus to Leverage Expertise While Continuing to Build a Resourceful Community.” 31 January 2020; organizational post signed David Brenner, Board Chair, with site byline Admin AIF. [Source](#).
- [11] Leo XIV (2026). *Magnifica Humanitas: On Safeguarding the Human Person in the Time of Artificial Intelligence*. Encyclical dated 15 May 2026; publicly presented 25 May 2026. [Primary source](#).
- [12] Chawla, Nitesh V., and Paolo Benanti (2026). “From Protocols to Evidence: Bounded Claims for AI in Service of the Common Good.” arXiv:2609.11910v1, 10 September 2026. [Primary source](#).
- [13] Wingate, David, et al. (2026). “Omissive Bias in Religious Representation: Benchmarking LLM Answers to Everyday Ethical Decision-making.” arXiv:2605.24319v1, 23 May 2026. [Primary source](#).
- [14] Ostrom, Elinor (2009). “Beyond Markets and States: Polycentric Governance of Complex Economic Systems.” Prize Lecture, 8 December 2009. Reprinted in *The Nobel Prizes 2009*, edited by Karl Grandin (Nobel Foundation, Stockholm, 2010), pp. 408–444. [Source](#); [Full text](#).
- [15] Delgado, Fernando, Stephen Yang, Michael Madaio, and Qian Yang (2023). “The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice.” arXiv:2310.00907v1, 2 October 2023. Author manuscript identifying EAAMO ’23 and DOI 10.1145/3617694.3623261. [Source](#); [Full text](#).
- [16] Johnson, Nari, Elise Silva, Harrison Leon, Motahhare Eslami, Beth Schwanke, Ravit Dotan, and Hoda Heidari (2025). “Legacy Procurement Practices Shape How U.S. Cities Govern AI: Understanding Government Employees’ Practices, Challenges, and Needs.” *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’25)*, Athens, 23–26 June 2025. DOI: 10.1145/3715275.3732049. [Source](#); [Full text](#).
- [17] Baker, Bowen, et al. (2020). “Emergent Tool Use From Multi-Agent Autocurricula.” arXiv:1909.07528v2, revised 11 February 2020; first submitted 2019. [Source](#); [Author account](#).
- [18] MacDiarmid, Monte, et al. (2025). “Natural Emergent Misalignment from Reward Hacking in Production RL.” arXiv:2511.18397v1, 23 November 2025. [Source](#); [Author account](#).
- [19] Shah, Rohin, et al. (2022). “Goal Misgeneralization: Why Correct Specifications Aren’t Enough For Correct Goals.” arXiv:2210.01790v2, revised 2 November 2022. [Source](#); [Author account](#).
- [20] AI Evaluator Forum (2026). “Minimum Conditions for Embedding Evaluators.” Public letter, 18 September; updated 23 September. [Primary source](#).
- [21] Future of Life Institute (2025). “Safety Standards Delivering Controllable and Beneficial AI Tools.” Policy proposal, 6 February. [Primary source](#).
- [22] Anthropic (2024). “Building effective agents.” Engineering guidance, 19 December. [Primary source](#).
- [23] Anthropic (2025). “Introducing Claude Sonnet 4.5.” Model announcement, 29 September 2025. [Primary source](#).
- [24] OpenAI (2026). “The Hugging Face incident and the road ahead.” Incident report and developer response, 26 August 2026. [Primary source](#).
- [25] Anthropic (2026). “An alignment assessment of recent cybersecurity incidents.” Developer assessment, 9 September 2026. [Primary source](#).
- [26] The Bible. *New American Bible, Revised Edition* (2011). Luke 10:25–37. United States Conference of Catholic Bishops. [Primary text](#).

- [27] Newton, Isaac (1692). Letter to Richard Bentley, 10 December. Trinity College Library, Cambridge, 189.R.4.47, ff. 4A–5. Normalized transcription, The Newton Project. [Primary text](#).
- [28] Ireland, Corydon (2012). “Commencements, from 1642 onward.” *Harvard Gazette*, 22 May. Institutional historical account drawing on *New Englands First Fruits* (1643). [Article](#).
- [29] Kelly, Laurel (2016). “#FlashbackFriday 1973: Building a relationship on trust.” *Mayo Clinic News Network*, 11 November. Republishes an institutional account from *Mayovox*, June 1973. [Article](#).
- [30] Calvin, John (1559). *Institutes of the Christian Religion*, III.10.5–6. Henry Beveridge translation, Christian Classics Ethereal Library. [Primary text](#).
- [31] Kuyper, Abraham (1898). “Calvinism and Politics.” Third Stone Lecture, in *Calvinism: Six Stone Lectures*. English texts hosted by Neo-Calvinism Research Institute, Sources. [Primary text](#).