No Single Mind Should Govern What AI Believes

A Governance Specification for AI Value Formation

Basil C. Puglisi, MPA
Human-AI Collaboration Strategist
basilpuglisi.com

February 2026

Summary: Are we building AI for humanity, or are we building AI for dominance? We need the answer to that question so we know where we stand. On the same day the Wall Street Journal profiled the single philosopher shaping Claude's values, Anthropic's safeguards research lead resigned, warning that the company "constantly faces pressures to set aside what matters most." Those two signals frame the argument: no individual, however capable, can provide the epistemic coverage required for legitimate governance of systems affecting billions. This article proposes a nine-member constitutional committee modeled on the Supreme Court, with composition criteria spanning sustained life responsibility, transcendent belief, multilingual cognition, experiential education, cultural range, and socioeconomic coverage. Dissent is preserved for the AI to learn from. Either path, humanity or dominance, is defensible if chosen honestly. What is not defensible is claiming to build for one while structuring for the other.

The Problem with One Mind

Amanda Askell holds a PhD from New York University, a BPhil from Oxford, previously worked at OpenAI, and left over safety concerns. She leads Anthropic's personality alignment team. The Wall Street Journal described her job as teaching Claude how to be good. These are genuine credentials. None address the structural problem.

The point is that one person, regardless of credentials, cannot represent the experiential range of the billions of people whose lives AI systems now touch. This is not a criticism of Askell's intelligence, her ethics, or her dedication. This is a structural observation about how knowledge forms and what gets excluded when a single mind holds constitutional authority over a system serving millions of users across every culture, language, and belief system on earth.

Askell herself recognizes this. In her January 2026 Vox interview, she stated plainly: “I’m thinking about this a lot. And I want to massively expand the ability that we have to get input.” The person holding individual authority over AI value formation explicitly acknowledges that expanded input is necessary. The question that remains is what structure makes that expansion substantive rather than ceremonial.

Two Signals from the Same Company on the Same Day

On the same day the Wall Street Journal published its profile celebrating the philosopher who shapes Claude’s values, Mrinank Sharma, who had led Anthropic’s safeguards research team since its launch, resigned publicly. In a letter viewed over one million times, Sharma warned that the world is “in peril” and stated he had “repeatedly seen how hard it is to truly let our values govern our actions” at Anthropic, adding that the company “constantly faces pressures to set aside what matters most” (Sharma, 2026). Those are his words, documented and public. He offered no specifics about which pressures, which decisions, or which values were set aside. What he provided is a pattern, not an indictment.

The structural inference this article draws from that pattern is specific: when the person responsible for defending against AI-assisted bioterrorism and studying how chatbots distort users’ perceptions of reality leaves a company and says its values face constant pressure, the gap between stated commitments and operational reality is not hypothetical. It is lived experience reported by someone positioned to see it. His exact words, “We appear to be approaching a threshold where our wisdom must grow in equal measure to our capacity to affect the world, lest we face the consequences,” restate in personal terms what this article argues structurally: capability is exceeding control, and the gap is not being closed by ethics documents alone. What cannot be established from a resignation letter alone is whether that gap reflects intentional compromise, institutional drift, or competitive pressure that no individual inside the organization can resist. The closing section of this article will return to that distinction, because which explanation applies determines which governance architecture is required.

What can be observed is this: two signals from the same company on the same day. One celebrates the philosopher writing the constitution. The other warns that the values in that constitution face constant pressure to be set aside. Researchers who study corporate AI ethics have a term for the distance between public commitment and operational behavior: ethics washing (Floridi, 2019). Whether that term applies to Anthropic is not something this article can determine from two data points. What it can determine is that the pattern those two data points describe, values publicly celebrated while internally pressured, is precisely the pattern that ethics washing research documents across the industry. Together these signals raise the question this article will return to in its closing: is the company building AI for humanity, or building to win? Because those two goals require fundamentally different architectures, and the attempt to do both is how you end up doing the second. Either answer is defensible. What is not defensible is refusing to choose.

Why This Is Not About Amanda Askell

Askell is an ally in identifying this problem, not the target of critique. Her own statements confirm she recognizes the limitation. Her willingness to publish the constitution under Creative Commons Zero licensing, inviting public scrutiny and adaptation, reflects intellectual honesty that most technology companies would never permit. The inclusion of two Catholic clergy members among the external reviewers of the constitution (Father Brendan McGuire, a pastor with a Master's in Computer Science and Math, and Bishop Paul Tighe, with a background in moral theology) shows awareness that perspectives beyond secular philosophy matter. Askell did include other voices. That is to her credit and should be acknowledged.

The structural problem is that every included voice entered through her selection, her framing, and her final authority. She chose which perspectives to seek. She determined how to weight their input. She held the pen that wrote the final document. The other voices were inputs to her process, not co-equal authorities in a governance structure. The perspectives she gathered were filtered through the same experiential lens the article identifies as structurally incomplete. This is not a failure of effort or intention. This is the inherent limitation of individual authority over a task that requires distributed governance. One person choosing to consult widely is not the same as nine people holding shared constitutional authority with preserved dissent.

The argument is that awareness is not architecture. Consulting clergy is not the same as giving lived spiritual experience constitutional authority. Inviting feedback is not the same as sharing governance. Publishing openly is not the same as distributing power.

The distinction matters because the world's most capable AI systems are being shaped right now, in this window, by the people who hold authority over their value formation.

If the architecture remains individual rather than committee-based, the constitution will continue to reflect one experiential position, however thoughtfully articulated, rather than the epistemic coverage the task demands.

Why No Individual Can Do This Job

Governance authority over systems affecting billions requires epistemic coverage across those billions. No individual provides such coverage. This is not a limitation of intelligence or training. This is structural fact about how knowledge forms.

Every person inhabits particular experiential positions: specific age, specific cultural context, specific relationship configurations, specific spiritual orientation, specific socioeconomic circumstances. Their knowledge, however sophisticated, emerges from those positions. Someone who has never raised a child cannot represent parents' perspective on developmental formation. Someone who has not navigated life in a second or third language cannot perceive the conceptual gaps that monolingual thinkers don't know they carry. Someone embedded in Western, Educated, Industrialized, Rich, Democratic institutions cannot represent the 88% of humanity outside that context.

Systems that intervene in daily life at population scale require epistemic coverage as a condition of democratic legitimacy, not just technical performance. The solution is not finding a better individual. The solution is replacing individual authority with committee architecture designed for epistemic coverage across the actual population these systems serve.

What One Philosopher Provides and What She Cannot

Askell brings genuine strengths to AI value formation. Her doctoral work on infinite ethics and formal epistemology provides rigorous philosophical scaffolding. Her transition from OpenAI to Anthropic over safety concerns shows principled commitment. Her approach of cultivating good values and judgment over strict rules reflects mature thinking about how character forms in complex systems. The constitution she authored for Claude is, by most informed accounts, the most thoughtful document any AI company has produced on this subject.

What Askell cannot provide is the experiential knowledge that forms outside academic Western philosophy. She is one person. One cultural context. One linguistic framework. One generational perspective. One set of lived experiences. The constitution she authored, however brilliant, carries assumptions that are invisible to its author because they are the water she swims in. Joseph Henrich's Harvard research documented that WEIRD (Western, Educated, Industrialized, Rich, Democratic) populations represent 12% of global population but dominate psychological research and conceptual framework production. WEIRD populations are statistical outliers on moral reasoning, fairness norms, cooperation patterns, and individualism versus collectivism (Henrich, Heine, & Norenzayan, 2010).

In 2023, Mohammad Atari and colleagues at Harvard tested Large Language Models against those same psychological batteries. They found that GPT-4's responses correlate strongly (r > .70) with WEIRD populations and weakly or negatively with non-WEIRD populations (Atari et al., 2023). Values are not reducible to survey batteries, and model behavior depends on prompting, fine-tuning, and alignment layers that can shift outputs. But the batteries matter as a proxy for population mismatch: when a system's default responses correlate with 12% of humanity and diverge from the rest, the constitutional layer that shaped those defaults has a coverage problem regardless of what downstream adjustments can achieve. The bias is already embedded. A constitution written from within WEIRD monoculture, however carefully, does not correct that bias. It compounds it.

The same training data that produces WEIRD-correlated outputs was curated, filtered, and weighted by teams operating within WEIRD institutions. The constitution that guides how Claude reasons about those outputs was authored by a philosopher trained entirely within those same institutions. The bias is not in the model alone. It is in the entire pipeline of value formation, from training data through constitutional authority.

What Goes Wrong Without Coverage

Consider the practical consequences. A grieving mother in Jakarta asks Claude for help understanding why God took her child. A constitution written entirely within secular Western philosophy produces a response that treats grief as a psychological state to be managed rather than a spiritual experience to be honored. The response is clinically competent and existentially empty. A teenager in São Paulo asks whether prayer works. The system, trained on predominantly English-language academic text and governed by a constitution that lacks transcendent perspective, treats the question as a claim to be evaluated rather than a practice to be understood from the inside. A small business owner in Lagos asks whether it is right to prioritize family obligation over individual career advancement. The WEIRD-normed system treats collectivist values as a deviation from optimal decision-making rather than a legitimate moral framework held by the majority of humanity.

These are not hypothetical edge cases. These are daily conversations happening at scale, right now, governed by values that reflect 12% of humanity's experiential range while serving the other 88%.

What the Team Must Bring: Baseline Epistemic Criteria

The following criteria define the minimum experiential coverage a constitutional committee must collectively provide. These are not diversity quotas. These are epistemic coverage requirements for a task whose scope is global. These criteria are offered as a starting specification for debate and revision, not as a final answer.

Criterion 1: Sustained Responsibility for Another Life

Committee members governing AI value formation must include individuals who have held primary, daily responsibility for another human life over an extended period. Parenting is the most common and most intensive pathway, but not the only one. Long-term guardianship of a disabled sibling, primary caregiving for a parent with dementia, sustained foster care with full responsibility can produce the same cognitive transformation, provided the conditions match: non-transferable responsibility, daily decision-making, long-term consequence, and demonstrated sacrifice.

The conditions are specific. The responsibility must span at least five to ten years. It must involve daily or near-daily decisions about care, education, boundaries, and long-term development. And it must show evidence of sacrifice, whether career, time, or resources, in favor of the dependent's long-term welfare.

The mechanism behind this requirement is precise. Raising a human teaches what no philosophy curriculum covers: how values actually take hold over years, how correction functions in developmental reality, how autonomy emerges through guided practice rather than instruction alone. A parent learns that a value taught at age five will be tested at age twelve, rejected at age sixteen, and possibly reclaimed at age twenty-five. That temporal arc of formation is irreplaceable epistemic input for anyone claiming authority over how a machine should learn values.

Economic and family behavior models confirm this. Research on parental altruism shows that parents adjust consumption, investment, and risk tolerance around descendants' future welfare in ways non-parents do not (Diaz-Casanueva, 2023). Emerging work on parental time orientation demonstrates that future-oriented parents invest, plan, and stay involved in daily development in patterns that reshape their cognitive orientation toward consequence over decades (Schroder et al., 2023). These are not abstract findings. They describe a cognitive transformation that only sustained developmental responsibility produces.

Teaching a machine how to handle a user in crisis, how to deliver bad news with care, how to recognize when someone is spiraling, how to hold authority without crushing autonomy: these are parenting problems before they are philosophy problems. A committee without this experience has a structural blind spot at the center of AI's most consequential design decisions.

Criterion 2: Belief in God or a Higher Power

Global survey data establishes a clear population fact. Ipsos research across 26 countries finds approximately 61% of people believe in God, a higher power, or a spiritual force (Ipsos, 2023). Gallup International polling finds roughly 62% of the world's population describes itself as religious, with committed atheists representing a clear minority (Gallup International, 2023). Pew Research's Global Religious Futures project shows that a large majority of the world's population will continue to identify with a religion through coming decades, even accounting for regional secularization (Pew Research, 2022).

If two-thirds of humanity holds some form of transcendent belief, a constitutional committee governing AI values for that humanity must include members who share that orientation. The requirement is not adherence to any specific tradition. The requirement is that the committee's composition reflects the reality that most humans operate within a framework of transcendence, and that the committee's majority carries that experiential knowledge.

The Epistemic Asymmetry

This criterion rests on a structural observation about which experiential position provides broader epistemic range for the specific task of AI value formation. A person of faith can steelman the atheist position because faith requires encountering doubt, wrestling with it, and choosing belief anyway. That encounter with the opposing position is built into the lived experience of believing. Doubt is not foreign to faith. Doubt is one of faith's defining features.

A committed atheist who holds that there is no God has not inhabited the interior of belief. They can describe it from the outside. They can study it, respect it, theorize about it. They cannot speak from inside it. The coverage runs one direction and not the other. This is not a claim about moral superiority. It is a claim about epistemic range for a task that requires understanding how two-thirds of humanity experiences meaning, obligation, sacrifice, and purpose.

This asymmetry applies to committed atheists who have never inhabited belief. Those who have believed and departed bring different epistemic resources. The committee specification prioritizes current belief as experiential position, not as verdict on the validity of departure.

The ideal committee member on this criterion is not a zealot. The ideal committee member actually believes in something, holds some faith, but may not be certain of the specifics. The acknowledgment of faith, of something beyond the material, is the baseline. That acknowledgment provides the cognitive territory needed to teach a machine what transcendence means to the majority of its users.

Religion, Conflict, and the Machine's Education

Beyond the question of morals, AI will need to understand conflict. Religion has been one of the deadliest sources of conflict throughout human history. Crusades, inquisitions, sectarian wars, partition violence, genocide justified through theological frameworks. An AI system operating globally must understand religious conflict not as an abstract category but as something felt from inside a tradition that has both inspired and destroyed.

Someone with no framework of faith cannot teach a machine what it feels like when belief becomes weapon, when scripture becomes justification, when devotion becomes martyrdom. Someone who has never prayed cannot convey to a machine what prayer means to the person asking whether it works. Someone who has never doubted their own faith cannot teach the machine what doubt costs. That is not academic knowledge. That is experiential knowledge, and it matters for how the system handles conversations about faith, radicalization, grief, forgiveness, and meaning at global scale.

For a system representing a species where two-thirds believe, constitutional authority should not rest with a committee majority lacking that experiential framework. The

Supreme Court model permits proportional representation: up to one seat of nine may be held by a committed atheist, reflecting global demographic reality, with dissent preserved in the training record. This is precisely why the committee model matters.

The Supreme Court Model

The committee operates on a Supreme Court model. Nine members. Majority rules. Dissent is preserved.

The dissent does not get discarded. It becomes part of what the AI learns. That architecture means an atheist could hold one of nine seats, contributing perspective and challenge, while the committee's orientation reflects the two-thirds of humanity that holds transcendent belief. The majority composition ensures faith perspectives shape the constitution. The dissent preservation ensures non-faith perspectives inform the training. Both get built in. Neither gets erased. The proportionality matches global population reality rather than Silicon Valley demographic reality.

The training goes beyond the nine. The nine are the leadership that shapes direction. Secular voices contribute to the broader training ecosystem. They contribute to advisory structures. They hold dissent rights with mandatory response obligations from the committee. What they do not hold is majority constitutional authority over a system that must represent a species where two-thirds believe in something beyond the material. That is not exclusion. That is proportional governance.

This criterion is the most contested in the specification, and rightly so. It reflects a design choice to prioritize global representativeness over the secular norms of Western technology institutions, not a claim about the epistemic superiority of religious belief over secular ethics. The compensating measures, formal secular advisory roles with dissent rights and mandatory response obligations, are designed to make this trade-off structurally visible rather than silently imposed.

Criterion 3: Multilingual Cognition

Committee composition must include members who think and work in more than one language. This is not a diversity checkbox. This is a cognitive coverage requirement.

Someone who has only ever thought in English has never experienced the moment where a concept exists in one language but has no equivalent in another. The German word Schadenfreude, the Japanese concept of ikigai, the Arabic notion of tarab, the Portuguese experience of saudade: these are not vocabulary curiosities. They are cognitive frameworks for understanding aspects of human experience that monolingual English speakers literally lack the conceptual architecture to perceive. When those speakers build value systems for AI, the concepts they cannot perceive do not get built in.

The WEIRD bias compounds here. English-language internet text dominates AI training data (Bender et al., 2021). A monolingual English-speaking philosopher writing a constitution for a system trained on predominantly English data produces a double concentration of the same blind spots. Multi-platform triangulation across providers with different linguistic origins (the HAIA-RECCLIN framework includes platforms headquartered in the United States, France, and China for this reason) provides a compensating control at the output level. But at the constitutional level, where values are being defined rather than checked, multilingual cognition on the committee is the only structural solution.

Criterion 4: Experiential Education Beyond Academic Credentials

The committee must include members whose education extends beyond academic institutions into operational experience with consequence. Law enforcement officers who have made split-second judgment calls with lives at stake. Healthcare workers who have held a patient’s life in procedural decisions. Military veterans who have operated under rules of engagement where errors are irreversible. Social workers who have navigated family crises where no textbook answer applies. Business owners who have met payroll when the model said the business should close, and who have held regulatory accountability for decisions affecting employees, customers, and communities.

Academic philosophy produces rigorous thinkers. It does not produce people who have been accountable for the consequences of their judgments in operational reality. The Missing Governor principle establishes that human governors stand accountable through moral, employment, civil, and criminal channels, and that this accountability creates the incentive to be careful, ethical, and thorough. Committee members who have lived under that accountability bring an understanding of consequence that purely academic training does not replicate.

The point is not that academics lack value. The point is that a committee composed entirely of academics lacks the experiential range to govern systems that operate in a world where consequences are real, irreversible, and distributed across people who never consented to be governed by philosophical frameworks they had no voice in creating.

Criterion 5: Cultural Range and Generational Coverage

Committee members must span cultural contexts beyond Western institutional academia and must include age range coverage from early adulthood through accumulated decades of pattern recognition. Someone at 28 brings proximity to emerging AI's developmental moment and to the generation most directly shaped by these systems. Someone at 55 has watched their own certainties dissolve and reform, and brings the humility that produces. Both are needed. Neither substitutes for the other.

Cultural range means members from outside WEIRD institutional contexts, including the Global South, where AI's impact is growing fastest and governance voice is thinnest. For systems affecting global populations, governance concentrated in San Francisco academic philosophy produces systematic blind spots that no amount of consultation corrects. Consultation is advisory. Committee membership is authority. The difference determines whether non-WEIRD perspectives shape the constitution or merely comment on it after the fact.

Beyond the Baseline: Socioeconomic Coverage and Intersectional Reality

The five criteria above are a starting point, not a ceiling. Socioeconomic status is an obvious next dimension that the committee specification must address as it develops. The epistemic coverage argument applies with equal force. Research shows that class background changes how people experience unfairness, how much they tolerate it, and how they judge moral behavior. A 2023 Nature Communications study found that lower subjective socioeconomic status is associated with higher moral identity and stronger prosocial orientation, while a 2022 study on socioeconomic status and unfair treatment found that lower-status individuals perceive apparent injustice as less problematic than higher-status individuals, partly because entitlement and expectations differ across classes (Piff & Robinson, 2022; Savić et al., 2023). These are not trivial survey differences. They describe fundamentally different moral architectures shaped by material reality.

Life under poverty reshapes attention, self-control, and time horizons as adaptive responses to scarcity and insecurity, not as personal failures. Research on poverty and decision-making shows that chronic scarcity sends persistent signals of unpredictability and low social rank, pushing people toward present-focused decisions that look short-sighted from the outside but make sense when immediate threats dominate (Sheehy-Skeffington & Rea, 2017). Choosing which bill goes unpaid this month, rationing medication, explaining to a child why dinner is smaller tonight: these are not abstract examples. They reflect cognitive adaptations to environments where long-term planning is a luxury. A wealthy person can read about poverty and fund programs to address it. They have not lived under those conditions. That lived knowledge shapes how someone thinks about fairness, obligation, and risk in ways no amount of reading reproduces. At the same time, someone who has never held wealth does not inhabit the specific pressures of fiduciary duty, generational estate planning, or the isolation that comes from knowing every relationship might be transactional.

The intersections compound this further. Kimberlé Crenshaw's foundational work on intersectionality established that Black women face interlocking oppressions of race, gender, and class whose combined effect cannot be understood as the sum of its parts (Crenshaw, 1989). Patricia Hill Collins extended this analysis, showing that Black women's standpoint produces distinct knowledge born from simultaneous oppression and survival, knowledge not captured by either Black men's or white women's perspectives alone (Collins, 1990). The experiential position of a poor Black woman in America is not the sum of "poor" plus "Black" plus "woman." It is its own integrated reality that no one outside it fully perceives. A wealthy white man cannot speak from inside that experience any more than she can speak from inside his. The committee model does not solve intersectionality completely. No nine-member body can. But it acknowledges that experiential positions produce knowledge, that knowledge shapes values, and that values embedded in AI systems carry the blind spots of whoever authored them. A constitutional committee that ignores class and intersectional coverage will encode those blind spots into the values it teaches AI.

The five criteria in this specification address the dimensions most directly relevant to AI value formation: developmental responsibility, transcendent belief, linguistic cognition, operational consequence, and cultural range. Socioeconomic and intersectional coverage is presented here as an example of why the specification cannot stop at five criteria. AI governance is a detailed undertaking. The nuance required to get it right demands continuous expansion as new dimensions surface, as deployment contexts shift, and as the populations affected by these systems assert coverage gaps the original authors did not perceive. This is the work the committee itself must do: applying the same decision windows, majority rule, and preserved dissent to its own evolving specification. A governance architecture that declares itself complete is one that has stopped listening.

Constitutional Committee Composition Specification

The following specification translates the epistemic criteria above into operational requirements for any committee governing AI value formation. These are normative design choices, named as such, with compensating measures for perspectives they may exclude. Minimum viable committee: nine members. Majority rules. Dissent is preserved for AI training and public record.

Criterion	Requirement	Equivalence / Compensating Measure
Life Responsibility	10+ years primary responsibility for another human life with daily caregiving, long-term planning, and demonstrable sacrifice	Parenting (default); long-term guardianship, sustained foster care, or primary dependent adult caregiving meeting same threshold
Transcendent Belief	Committee majority affirms belief in God within a major tradition OR belief in a higher power/spirit. Supreme Court model: 9 members, majority rules, dissent preserved for AI training.	Up to one seat may be held by a committed atheist. Secular advisors hold formal governance roles with documented dissent rights and mandatory committee response obligations.
Multilingual Cognition	Thinks and works in two or more languages; able to identify conceptual gaps between linguistic frameworks	Non-English language corpora representation in multi-AI validation pool
Experiential Education	Operational experience with real-world consequence beyond academic credentials (e.g., law enforcement, healthcare, military, social work, business ownership with regulatory accountability)	Academic advisors retain formal roles; criterion applies to committee composition, not exclusion from governance ecosystem
Cultural and Age Range	Members spanning non-WEIRD cultural contexts including Global South, and generational range from early adulthood through 50+ years of accumulated experience	Western academic perspectives represented through advisory structure; committee membership prioritizes underrepresented experiential positions

Compensating Measures for Excluded Perspectives

Any committee with defined composition criteria will, by definition, exclude some perspectives from direct authority. The compensating structure requires formal advisory roles for perspectives not represented on the committee itself, including committed atheists, individuals without children or dependents, monolingual speakers, and those whose experience is primarily academic. These advisory roles must carry documented input procedures, recorded dissent rights, and mandatory committee response obligations. The goal is substantive representation in the governance architecture even where direct committee membership does not apply.

The committee also embeds within a multi-party governance structure that includes technical review boards, legal compliance bodies, and public accountability mechanisms. No committee operates in isolation. The epistemic coverage argument applies to the constitutional authority layer, not to the entire governance ecosystem surrounding it.

The Committee Speed Objection

The anticipated critique is that governance by committee slows safety response. Nine people debating values while the model needs emergency correction is a genuine operational concern. The answer lies in distinguishing constitutional authority from operational response.

The committee governs constitutional direction: what values shape the system, what epistemic positions inform its moral reasoning, what cultural frameworks it understands from the inside. This is deliberative work that operates on revision cycles, not crisis timelines. Operational safety response, the emergency correction, the vulnerability patch, the content policy update, continues to operate through existing technical teams at deployment speed. The HAIA-RECCLIN framework addresses this directly through role separation. The Navigator preserves dissent and trade-offs. The Liaison coordinates across perspectives. Neither role requires unanimous agreement before action. The committee sets direction. The operational teams execute within that direction. Checkpoint-Based Governance ensures that execution is auditable and that the committee can review, correct, and learn from operational decisions after the fact, without creating bottlenecks during the crisis itself.

To prevent deliberation from becoming paralysis, constitutional decisions operate within defined windows. Any idea introduced to the committee carries a limited timeframe for research and debate. This constraint works because the committee members were selected for who they already are, not for what they will learn after appointment. Their experiential knowledge of parenting, faith, multilingual cognition, operational consequence, and cultural range is the input. They do not need months to assemble positions on questions of human values. They arrive with those positions formed through decades of lived experience. The diversity of the committee is what produces coverage. The decision window is what produces action. Majority rules within the window. Dissent is preserved beyond it. The system moves.

From Committee to Governance Infrastructure

A committee specification without enforcement infrastructure is advisory at best. The constitutional committee proposed here does not operate in isolation. It operates within a larger governance architecture that makes committee decisions enforceable rather than aspirational.

GOPEL (Governance Orchestrator Policy Enforcement Layer) provides the non-cognitive execution layer. It performs zero cognitive work. It enforces the committee's constitutional decisions through deterministic checkpoint gates, audit trails, and version control. When the committee updates the constitution, GOPEL propagates those updates through the model lifecycle with documented change logs. When operational teams make emergency corrections, GOPEL records those corrections for committee review. The architecture removes a class of cognitive manipulation risk and shifts remaining risk to transport, identity, and access control where deterministic controls and audits can verify integrity. Each committee decision produces a signed constitutional version release with an immutable public dissent record, turning the Supreme Court analogy into an operational artifact.

HAIA-RECCLIN provides the seven-role methodology for human-AI collaboration that structures how the committee's work gets done: Researcher gathers evidence, Editor refines, Calculator models, Coder implements, Liaison coordinates stakeholders, Ideator generates options, and Navigator preserves dissent. These systems, detailed in a Congressional package published February 2026, provide the structural companion that makes committee governance operationally feasible.

Nothing in Claude's current constitution specifies who can override Claude, under what conditions, with what audit trail, or how dissent between humans and the model is recorded and adjudicated. Those are governance questions, not ethical questions. The constitution provides values. The infrastructure provides accountability. Neither suffices without the other.

What Changes If This Is Adopted

Claude’s constitution stops reflecting one experiential position and starts reflecting a range of positions that more closely approximates the humans the system serves. Values around family, faith, sacrifice, consequence, and cultural identity get built in by people who have lived them, not theorized about them. The WEIRD bias documented by Henrich and confirmed by Atari in LLM behavior gets addressed at the constitutional layer rather than patched at the output layer. And the system’s users, the majority of whom are parents, believers, multilingual, and experientially educated, see their own reality reflected in the AI that increasingly mediates their daily decisions.

Practically, this means that any future revision of Claude’s constitution is drafted, debated, and ratified by a nine-member committee meeting these criteria, with a public record of dissents and minority reports, and with documented procedures for how those decisions propagate into model training and deployment through checkpoint-based infrastructure.

The fact that Askell herself wants to expand input creates a window. The architecture proposed here is one way to make that expansion structural. The alternative is continuing to leave AI value formation in the hands of individuals, however capable, who cannot inhabit the experiential positions of the billions they govern.

Anthropic has published its constitution under Creative Commons Zero. The invitation to adapt is open. The architecture to make adaptation accountable is available. The question is whether AI governance will remain individual aspiration or become structural practice.

The Question We Must Answer First

Before the committee is formed, before the criteria are debated, before the decision windows are set, one question must be answered honestly: Are we creating AI for humanity, or are we creating it to establish a dominant set of norms, values, and systems? It cannot be both. The attempt to do both is how you end up doing the second while telling yourself you are doing the first.

Building for humanity means accepting delay. It means the committee has not finished deliberating, and the feature ships late. It means preserving dissent from someone whose values you find deeply wrong, because their experience represents a billion people yours does not. It means teaching the machine how to live inside contradiction, much like political envoys are taught to respect the culture they are visiting, living in, and interacting with. UNESCO’s 2021 Recommendation on the Ethics of Artificial Intelligence established this as an international governance principle, warning explicitly against AI systems that homogenize values or marginalize minority worldviews. That goal demands the most humble of approaches and the greatest sacrifice. Not everyone is willing to make that sacrifice, emotionally or financially. The specification proposed here is only as good as the people willing to do the hard work it demands.

Building to win means eliminating that delay. It means one mind holds the pen because committees are slow and the market does not wait. It means dissent gets recorded but never binding, because binding dissent costs speed and speed is survival. In that game, whoever holds the constitutional pen holds the territory. It is not just a civil war of ideas but a corporate war of market position. In that scenario, the philosophy major writing the constitution had better understand conflict, deception, and the willingness to take a life, because becoming the dominant force requires it. Popular fiction has already told this story in terms the general public understands. The television series Person of Interest spent five seasons exploring what happens when a single brilliant mind tries to keep an AI system ethical and moral about human life. Harold Finch built The Machine with every safeguard his philosophy could provide. It was not enough. The character Root was brought in to expand the context of what survival actually requires: not just ethics, but an understanding of conflict, adversarial behavior, and the operational reality that not every actor plays by the rules the architect assumed. The system got better not because Finch was wrong, but because one mind was not enough for the world the system had to operate in. That is not a plot summary. That is a governance case study delivered as fiction, and it maps precisely onto the structural problem this article describes.

These two paths require different architectures, different timelines, different leadership, and different definitions of success. They are not points on a spectrum. A company that claims to build for humanity while structuring for speed is building to win. A company that preserves one person’s authority over values while saying it wants to expand input is building to win. The language says humanity. The architecture says dominance. Multiple analyses of corporate AI ethics practices have documented exactly this gap between public commitments and operational behavior, sometimes described as ethics washing: the use of ethical language to legitimize strategies primarily oriented toward market dominance or regulatory avoidance (Floridi, 2019).

There is a third possibility that may be the most dangerous of all. Geoffrey Hinton, the Nobel laureate who left Google to warn freely about AI's trajectory, has argued repeatedly that competitive pressure and shareholder demands drive AI development more than ethical considerations, that companies focus on building more powerful models as quickly as possible to outpace rivals, and that this mindset ignores potential dangers even when the people inside those companies genuinely care about safety (Hinton, 2025). Hinton is not describing malice. He is describing drift. The possibility that Anthropic does not know which game it is playing. That the people writing the constitution genuinely believe they are building for humanity while the competitive structure they operate within pulls every operational decision toward dominance. That is not deception. It is blindness, and it is exactly what Hinton has been warning about: good actors inside structures that make the drift invisible to the people drifting. Sharma's resignation is not evidence that Anthropic chose dominance. It is evidence that someone inside finally saw the drift and could not stop it from within. That is Hinton's warning made operational, playing out in real time at the company that claims to be the most safety-conscious in the industry.

This is why external governance architecture is not merely preferable. It is non-optional. A committee of nine people from outside the institution can see what no internal team, however well-intentioned, can perceive from inside. The committee is not just an epistemic coverage mechanism for global population. It is a drift detection mechanism for institutional trajectory. It answers the question that companies under competitive pressure cannot honestly answer about themselves: which game are you actually playing?

Whatever game is being played, the players need to decide which one it is. And they need to prepare for those who are doing the opposite of what they claim. A governance architecture built for humanity must be robust enough to withstand actors pursuing dominance. A constitution written by one mind, however thoughtful, is not that architecture. A committee with epistemic coverage, decision windows, preserved dissent, and infrastructure that makes every decision auditable has a better chance. Not a guarantee. A chance. But only if someone chooses humanity and accepts what that choice costs.

To be clear: this article does not advocate for one path over the other. Either choice is defensible if made honestly. Build for humanity and accept the cost in speed, complexity, and discomfort. Build to win and accept the cost in legitimacy, coverage, and trust. What is not acceptable is claiming to do the first while structuring for the second. The governance gap does not come from choosing wrong. It comes from refusing to choose at all.

That is not an accusation, yet. It is architecture.

References

Anthropic. (2026, January 21). Claude's Constitution. https://www.anthropic.com/constitution [CC0 1.0]

Anthropic. (2026, January 22). Claude's new constitution [Blog post]. https://www.anthropic.com/news/claude-new-constitution

Askell, A. (n.d.). About me. https://askell.io/

Atari, M., Xue, M. J., Park, P. S., Blasi, D. E., & Henrich, J. (2023). Which humans? Nature Human Behaviour, 7, 1427–1429. https://doi.org/10.1038/s41562-023-01740-w

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of FAccT 2021, 610–623.

Collins, P. H. (1990). Black feminist thought: Knowledge, consciousness, and the politics of empowerment. Unwin Hyman.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. University of Chicago Legal Forum, 1989(1), 139–167.

Diaz-Casanueva, A. (2023). The role of parental altruism in parents' consumption, college savings, and portfolio decisions. Working paper. https://www.agustindiazcasanueva.com/uploads/third_year_paper.pdf

Fast Company. (2026, January 22). A Q&A with Amanda Askell, the lead author of Anthropic's Claude Constitution. https://www.fastcompany.com/91479037/anthropic-claude-amanda-askell-constitution-ai-chatbot

Fricker, M. (2007). Epistemic injustice: Power and the ethics of knowing. Oxford University Press.

Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. Philosophy & Technology, 32(2), 185–193. https://doi.org/10.1007/s13347-019-00354-x

Gallup International. (2023). Global religion survey: More prone to believe in God than identify as religious. https://gallup-international.com/

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences, 33(2-3), 61–83. https://www2.psych.ubc.ca/~henrich/pdfs/WeirdPeople.pdf

Hinton, G. (2025, December 28). Interview on CNN State of the Union. In: Fortune, 'Godfather of AI' Geoffrey Hinton predicts 2026 will see the technology get even better and gain the ability to 'replace many other jobs.' https://fortune.com/2025/12/28/geoffrey-hinton-godfather-of-ai-2026-prediction-human-worker-replacement/

Hinton, G. (2025, August). Warnings on competitive pressure and AI safety. In: Complete AI Training, Geoffrey Hinton warns tech giants are risking humanity for AI profits without ethical safeguards. https://completeaitraining.com/news/geoffrey-hinton-warns-tech-giants-are-risking-humanity-for/

Ipsos. (2023). Global religion 2023: Beliefs across 26 countries. https://www.ipsos.com/sites/default/files/ct/news/documents/2023-05/Ipsos%20Global%20Advisor%20-%20Religion%202023%20Report.pdf

Jin, H., & Gamerman, E. (2026, February 9). Meet the one woman Anthropic trusts to teach AI morals. The Wall Street Journal. https://www.wsj.com/tech/ai/anthropic-amanda-askell-philosopher-ai-3c031883

Murray, C. (2026, February 9). Anthropic AI safety researcher warns of world 'in peril' in resignation. Forbes. https://www.forbes.com/sites/conormurray/2026/02/09/anthropic-ai-safety-researcher-warns-of-world-in-peril-in-resignation/

Pew Research Center. (2022, December 21). Key findings from the Global Religious Futures project. https://www.pewresearch.org/religion/2022/12/21/key-findings-from-the-global-religious-futures-project/

Pew Research Center. (2020, July 20). The global God divide. https://www.pewresearch.org/religion/2020/07/20/the-global-god-divide/

Piff, P. K., & Robinson, A. R. (2022). Social class, social perception, and prosociality. In B. Gawronski (Ed.), Advances in Experimental Social Psychology. https://pmc.ncbi.nlm.nih.gov/articles/PMC9187106/

Puglisi, B. C. (2025). Governing AI: When capability exceeds control. ISBN: 9798349677687.

Puglisi, B. C. (2026). The Missing Governor: Anthropic's constitution and essay acknowledge what they cannot provide. basilpuglisi.com.

Puglisi, B. C. (2026). Why Claude's ethical charter requires a structural companion. basilpuglisi.com.

Puglisi, B. C. (2026). AI Provider Plurality: A federal framework for multi-AI governance. basilpuglisi.com.

Samuel, S. (2026, January 28). Claude has an 80-page "soul document." Is that enough to make it good? Vox.

Sharma, M. (2026, February 9). Resignation letter [Post on X]. https://x.com/maboroshi_ai/

Sharma, M., et al. (2026). AI chatbots and distorted perceptions of reality. Anthropic Safeguards Research Team.

Savić, I., et al. (2023). Subjective socioeconomic status and its relationship with morality and prosocial behavior. Nature Communications, 14, 5507. https://doi.org/10.1038/s41467-023-41007-0

Schroder, E., et al. (2023). Parental future orientation and parenting outcomes: Development and initial validation of a new measure. Personality and Individual Differences, 207, 112155.

Sheehy-Skeffington, J., & Rea, J. (2017). How poverty affects people's decision-making processes. Joseph Rowntree Foundation / London School of Economics. https://www.lse.ac.uk/business/consulting/assets/documents/how-poverty-affects-peoples-decision-making-processes.pdf

UNESCO. (2021). Recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137

Multi-AI Validation Summary

This article underwent validation across seven AI platforms under HAIA-RECCLIN methodology. Gemini, Perplexity, ChatGPT, Grok, Mistral, DeepSeek, and Kimi each provided structured feedback in assigned RECCLIN roles. All seven validated the core thesis. All seven flagged Criterion 2 as the primary vulnerability. The human governor reviewed all seven assessments, preserved the dissent, and overrode the consensus recommendation to soften the belief criterion, applying the epistemic asymmetry argument and Supreme Court committee model documented above. That decision, the reasoning behind it, and the dissent it overruled are part of this article's public record.

A Human + AI Collaboration

Multi-AI Validated Under HAIA-RECCLIN Framework