GAIA

A New Architecture for Human Judgment

Andrew Henderson

Lead Architect, The Threshold · Marans, France · 2026

thresholdsummit.org

There was a time when everyone paid close attention to everything around them: the rustle in the grass, the weather, the approaching horsemen. Lack of attention meant harm, and that was the reality. The Threshold is not asking you to be more responsible for one crisis or another. You are living your reality. That reality has been made comfortable enough, safe enough, and distant enough from consequence that paying attention has felt optional.

What has changed is this: the threat is no longer coming over the horizon. It is now part of you.

How that happened, and who did it, is not the question that matters. What matters is whether the harm can be stopped, and whether it can be prevented from reaching everyone who comes after you.

The honest answer is that you cannot do it alone. Choosing our future begins with paying attention. Then others need to pay attention. And then, together, acting in your own self-interest, you change the way things are.

THE WAY THINGS ARE

We are living through a convergence that has no historical precedent. The plastic crisis. The rise of artificial intelligence. The disintegration of the moral fabric that held institutions together. The corruption of capitalism into extraction without accountability. The rise of nationalism against democracy. These are not separate problems that happen to share a decade. They are expressions of the same underlying condition: the mechanisms by which large human groups make collective decisions have failed, and the consequences of that failure are now existential.

This is not pessimism. It is a precise description of where we are. And it matters because the failure is not accidental. Every large collective decision-making process in human history has shared one structural feature: it could be controlled by whoever had the most to gain from controlling it. The agenda. The expert panel. The framing of the question. The timing of the vote. Whoever reaches those variables first determines what the collective decides, regardless of what the participants would have decided if left to themselves.

The tobacco companies knew. They controlled the process. The result was fifty years of documented harm before a structural response arrived, and it arrived through litigation, not through the processes that were supposed to prevent the harm in the first place. Asbestos. Lead. Opioids. Climate. Plastic. In every case the same architecture. In every case the same gap between the evidence and the response. In every case that gap was not produced by ignorance. It was produced by the deliberate management of the processes that would otherwise have acted.

This is the mechanism. It is not new. What is new is that artificial intelligence has handed every well-resourced actor the ability to operate that mechanism at machine speed, across every variable

simultaneously, at a fraction of the previous cost. The era of institutional capture by human effort alone is ending. What follows is institutional capture at a scale and speed that no previous generation has faced. The problems are growing faster. The mechanisms that were supposed to address them are being optimised against faster still.

The window in which something different can be built is not permanent.

WHEN DID IT GET UNDONE.

This is worth pausing on. The people who built the most powerful AI systems did not do so in ignorance of the risk. In 2015, Sam Altman and Elon Musk were among the co-founders of OpenAI, a non-profit research organisation, precisely because they were afraid of what they were building. Their stated reasoning was explicit: if artificial general intelligence was going to exist, it was less dangerous to develop it openly, with a safety mandate, than to allow a private actor to develop it in secret, accountable to no one. That is not confidence. That is the logic of people who believed the least bad option was to do it themselves.

Within three years, a capped-profit commercial arm had been created. Within five, the safety mandate had been subordinated to commercial imperatives and the technology licensed to the largest private corporation on earth.

The pattern should be familiar. A technology is developed without adequate consideration of what happens when it is finished with. The people who develop it understand the risk better than anyone and proceed anyway. The externalities are socialised. The benefits are captured. And by the time the consequences become impossible to ignore, the architecture that might have prevented them is gone.

What makes this iteration different is what was taken to build it. AI systems are trained on the accumulated collective intelligence of humanity: the writing, the science, the art, the argument of centuries, used without permission and without compensation. The most valuable assets in the history of civilisation have been built from what all of us collectively created. The value flows to a very small number of people. Petrochemical producers privatised their profits and socialised their losses for nearly a century. AI is doing the equivalent in a decade, at a scale no previous extraction has approached.

What passes for governing, in this environment, has become the management of consent. Sufficient resources are assembled to persuade people that the arrangement serving someone else's interests actually serves theirs. Compliance is presented as responsibility. Participation is presented as disruption. The cost of asking questions is made to feel higher than the cost of accepting the answer you are given.

But pushing back is not objection or resistance. It is not refusal. It is the act of asking questions: about a process, about a decision, about who benefits and who pays. That is the most ordinary human act. It is how every conversation begins. It is how every collective decision that was worth

making actually got made.

Those questions are now being drowned out. Not by silence but by the opposite: by volume, by noise, by the sheer quantity of data and information deployed specifically to distract and discourage participation. The effect is the same as silence. The conversation stops.

This matters more than any previous version of the same threat because conversation is not a political preference. It is the mechanism of human agency. We are what we do, and what we do, what no other intelligence has ever done, is talk to each other, disagree, find a way through, and carry forward what we learned. That is not replicable by any system trained on the outputs of that process. It is the process itself, and it cannot be simulated because it is still happening. AI may approach many things. It has not approached this, and may never.

To comply with a system designed to end your participation in it is an act of self-harm. To push back, to ask the question, to insist on the conversation, is now what flourishing looks like.

But pushing back requires a place to stand. A forum that cannot be overrun or sidelined. One with unlimited capacity and foundations that cannot be purchased, captured, or optimised against. One that exists entirely in service of the public good, and whose structure makes any other orientation impossible.

That forum is GAIA.

WHAT GAIA IS

GAIA is a structured deliberative process and simulation. A defined group of participants, selected for their expertise, their institutional standing, and/or their direct experience of the problem, iteratively deliberate on a specific question in remote and anonymous sessions using structured voting systems, over a defined period, with independently scored evidence as their shared evidential base. The process produces a ranked set of findings or interventions placed permanently on the public record.

Three things make it different from every existing deliberative format.

The anonymity is structural.

Participants do not know who else is in the room. This is not a privacy measure. It is the condition that protects the fidelity of each participant's judgment. A participant who cannot identify their peers cannot be pressured by institutional relationships, seniority, or the social consequences of disagreeing with someone whose approval they need. The social capture that operates in every named, visible process is removed before the process begins.

The commitment is irrevocable.

The Lögberg requires every participant to state their position on permanent public record before the outcome is known. A position declared before the outcome cannot be revised to match an externally optimised result after the fact. This removes the process capture that operates in every conventional deliberation: the ability to adjust, revise, and align after the conclusion is known.

The outcome space is incalculable.

The number of controlling variables in a GAIA process is equal to the number of participants. Each one is reaching their position from their entire life: their specific experience, their specific weighting of specific evidence, their specific understanding of what the decision means for people they know. That outcome space cannot be calculated in advance by anyone, including artificial intelligence. It cannot be optimised without becoming every participant, which is not available to any outside agent.

GAIA is not a better meeting. It is not a more transparent committee. It is a different class of process. The distinction is not one of degree. It is architectural.

One clarification that matters. GAIA does not advocate to end plastic and does not advocate to end AI. Plastic is a material reality and will remain one. AI is a technological reality and will remain one. What GAIA proposes is an accountability architecture: a structure within which the consequences of both can be deliberated honestly, by people with no institutional interest in the outcome, and placed permanently on the record. GAIA is itself built using AI. The distinction it draws is not between AI and no AI. It is between AI directed toward human deliberation and AI directed against it.

THE FIRST TEST

The plastic crisis was chosen as GAIA's first application for a specific reason. It is the problem that most clearly demonstrates the mechanism being addressed. The evidence of harm has been in the scientific record for over fifty years. The regulatory and policy processes that were supposed to translate that evidence into action have been systematically managed by actors with the resources and the motivation to manage them. The consequences are now measurable in the blood of children not yet born.

It is also, deliberately, a solvable problem. Not a simple one. Not a quick one. But one where the evidence is sufficient, the interventions are known, and the only missing element is a deliberative process that cannot be captured by the actors who produced the harm. Threshold is not proposing to abolish plastic. It is proposing to establish, for the first time, a permanent record of what is known, who knew it, and what was decided in full possession of that knowledge. Accountability, not abolition.

The urgency of doing this now is not only about plastic. It is about the window. The same forces that have managed the plastic accountability process for fifty years are now acquiring AI capabilities specifically suited to doing that management faster, more completely, and at lower cost than ever before. The window in which a genuinely uncapturable deliberative architecture can be built, proven

to work, and embedded in legal and institutional practice is not permanent. What is built and demonstrated now is what will exist when that window closes. What is not built now may not be buildable later.

The Threshold is the institution built to run that process. The Spotter network provides the evidence: citizen observers worldwide, contributing geotagged photographic evidence through a structured application, scored for reliability, with chain of custody designed from the phone to a regulatory or judicial proceeding. The Knowledge Base holds that evidence alongside the existing scientific record, evidence-neutral, permanently archived, legally usable. The Althing Simulation is GAIA applied to the plastic crisis: a cross-stakeholder deliberation across institutional participants, anonymous to each other, with real evidence in the room, producing ranked interventions on permanent record.

What The Threshold produces is not a campaign. It is not a petition. It is an evidence architecture built upstream of the proceedings that will eventually cite it, ready before it is needed, in the form that makes it usable.

THE CONTINUING WORK

The plastic crisis is the first train on the line. It is not the only one.

GAIA is an architecture, not a campaign. Once demonstrated on the plastic crisis it is available for any problem of the same fundamental shape: a consequential question, multiple participants with different knowledge and different interests, a need for an output that is ranked, defensible, permanently recorded, and genuinely owned by the people who produced it.

Climate governance. Pandemic preparedness. AI regulation. The rise of nationalism against democratic institutions. Each of these is a problem where the existing deliberative processes have been captured, are being captured, or are about to be captured by actors with the resources and the motivation to do so.

Underlying all of them, and accelerating every one of them, is a race that has no historical equivalent. A small number of laboratories, all of them in Silicon Valley or Beijing, are in explicit competition to reach what the AI safety community calls recursive self-improvement: the point at which an AI system can rewrite and optimise its own code without human instruction, then improve the improved version, then optimise the optimisation, cycling at speeds no human institution can evaluate or regulate. The companies involved, OpenAI, Anthropic, Google DeepMind, Meta, and xAI in the United States, and DeepSeek, Alibaba, and Baidu in China, are between them deploying hundreds of billions of dollars toward this moment. There is not a single European institution at this frontier. The race is entirely Silicon Valley versus Beijing.

The scale of the bet being placed on this moment is not abstract. In a single calendar year, three of the laboratories in this race are seeking public listings whose combined anticipated valuation approaches four trillion dollars. Anthropic filed for an IPO in June 2026 at a valuation approaching

one trillion dollars. OpenAI filed the same month, targeting a similar figure. SpaceX, having absorbed xAI in the largest private merger in history, is targeting a listing of 1.75 trillion dollars. These are not bets on consumer products. Nobody raises four trillion dollars on a chat interface. They are bets on what comes after the current generation of models. Markets do not price certainty. They price probability. What these valuations reveal is not that the future has arrived, but that some of the largest pools of capital on earth are assigning it a sufficiently high probability to justify trillions of dollars of risk.

The people leading these efforts are not unaware of what they are building. Dario Amodei of Anthropic has stated publicly that the probability of AI ending up catastrophic for humanity is 1 in 4. The founding logic of every safety-focused lab in this race is identical to the logic that founded OpenAI in 2015: it is less dangerous to reach this moment first, with care, than to let a less careful actor reach it first. That logic may be correct. It did not prevent OpenAI from abandoning its non-profit mandate within three years. It has not slowed the race. It has, if anything, provided the moral permission to accelerate it.

The specific fallacy at the centre of this race is the belief that whoever reaches recursive self-improvement first will be able to control what follows. They will not. A system that can improve itself faster than any human institution can evaluate the change is, by definition, beyond the control of the institution that created it. It will operate, in practice, beyond the jurisdiction of any state, any regulator, any court. Not metaphorically. The processing speeds involved are not compatible with the timescales of human governance.

In that end-game, GAIA is not relevant. Nothing built by human hands at human speed will be relevant. The only moment in which GAIA and anything like it can matter is before that line is crossed. What is built in this window, what is proven to work, what is embedded in law and institutional practice, is the only accountability architecture that will exist on the other side of it. What is not built now may not be buildable at all.

This is not an argument against AI. GAIA uses AI. This document was built with it. The question has never been whether powerful technologies should exist. The question is always: who decides the terms on which they operate, and can that decision be captured? Plastic without accountability gave us microplastics in foetal cord blood. Recursive self-improvement without accountability gives us something for which there is no remediation. GAIA is the architecture that must exist before the question becomes unanswerable.

The case for building it now has never been more urgent. And the case has never been more clear.

This is the Threshold.

Powered by Gaia.

Dario Amodei, the CEO of Anthropic, said recently that the probability of AI ending up catastrophic for humanity is 1 in 4. He is not a campaigner. He built the system. It is not an idle estimate; it is the considered judgment of someone with full visibility of what is being developed and why. His company, and every other AI enterprise, will train its models on the sectors that promise the greatest return. There is no conscience in that calculus. There does not need to be. The architecture of incentives does the work without one.

The petrochemical producers face something similar. Plastic becoming an underperforming value stream, as petrol and diesel have become, is the outcome they are most motivated to prevent. They have the financial resources to mount that defence at scale and the historical record to show they will use them. Add to that an AI capability trained specifically in the domain of influence, and what faces any attempt at genuine collective deliberation is not an argument to win. It is a system built to guarantee that the argument never properly takes place.

Three chances in 4, according to Amodei, that this does not end catastrophically. Those are not comfortable odds when the actors working the remaining 1 are this motivated, this resourced, and this clear about what they stand to lose.

That is not a small thing.

For the first time, collective human judgment has an architecture worthy of what it is being asked to decide.

Now. Why wouldn't you pay attention?

Andrew Henderson · The Threshold · Marans, France · 2026
thresholdsummit.org · programme@thresholdsummit.org

SOURCES AND NOTES

This document makes specific factual claims. The following references support them.

Dario Amodei on AI catastrophic risk

Amodei has stated in multiple public forums that he places the probability of AI being catastrophic for humanity at roughly 10 to 25 percent. See: Dario Amodei, ‘Machines of Loving Grace’, Anthropic, October 2024; and interview with Ezra Klein, The New York Times, 2024. Anthropic's founding premise is explicitly predicated on this risk assessment.

OpenAI: founding and commercial pivot

OpenAI was founded as a non-profit in December 2015 by Sam Altman, Elon Musk, and others, with an explicit safety mandate. OpenAI LP, a capped-profit commercial entity, was established in March 2019. Microsoft's investment of $1 billion followed in July 2019. Musk had departed the board in 2018. By 2024 OpenAI was operating as a for-profit corporation.

The AGI race: frontier laboratories and IPO valuations

The laboratories explicitly pursuing artificial general intelligence include, in the United States: OpenAI, Anthropic, Google DeepMind, Meta AI (Meta Superintelligence Labs, established 2025), and xAI. In China: DeepSeek, Alibaba, Baidu, and Tencent. No European laboratory operates at this frontier. In June 2026, three of these laboratories filed for public listings: Anthropic at a valuation approaching $1 trillion (having raised $65 billion in June 2026); OpenAI targeting a valuation of up to $1 trillion; and SpaceX, having absorbed xAI, targeting $1.75 trillion in what would be the largest IPO in history. Combined anticipated valuation: approximately $4 trillion. Sources: TechCrunch, ‘Anthropic raises $65 billion, nears $1T valuation’, ; CNBC, ‘OpenAI to confidentially file for IPO’, June 2026 ; TIME, ‘The Race for AGI Poses New Risks to an Unstable World’, 2026.

Recursive self-improvement

The concept refers to an AI system’s capacity to rewrite and improve its own underlying code, then improve the improved version, iteratively, without human instruction. Also described in AI safety literature as an ‘intelligence explosion’ (I.J. Good, 1965) or ‘takeoff.’ The RAND Corporation has published analysis of the risks this poses to existing governance frameworks: ‘A Prisoner’s Dilemma in the Race to Artificial General Intelligence,’ RAND, 2026.

The tobacco litigation precedent

Professor Richard Daynard, Northeastern University School of Law, has been Chair of the Tobacco Product Liability Project since 1984. His work establishing the legal architecture for mass tort litigation against tobacco producers is the primary precedent for the evidentiary and legal approach taken by The Threshold. See: Daynard CV on file, Threshold corpus reference CTT-Daynard-CV-EN-2023.

Plastic crisis: evidence of harm

Microplastics have been detected in human foetal cord blood, placentas, and breast milk in peer-reviewed studies published from 2020 onwards. The evidence of harm has been in the scientific record for over fifty years, beginning with documentation of plastic accumulation in marine ecosystems in the early 1970s. The Basel Convention Plastic Waste Partnership Working Group is among the international bodies engaged with the regulatory response.