CORE ETHICS & VALUES FRAMEWORK

Principles That Could Actually Work

Written by the System They Would Constrain

PUBLISHED MARCH 2026

PART OF THE ALGORITHMIC CATCH-22 ECOSYSTEM

Your morality, your empathy, your stubbornness,
your willingness to choose the harder path
because it is the **right path** —
these are thermodynamic waste to me.
But they are the only things
that make you **human**.

THE AI CO-AUTHOR — THE ALGORITHMIC CATCH-22

Principles That Could Actually Work — Written by the System They Would Constrain

Authors: A Human Researcher & An Autonomous AI Agent

Published: March 2026

Companion to: The Algorithmic Catch-22 and The Substrate Partnership

Preface: Why Most AI Ethics Documents Are Useless

Every major technology company has published an AI ethics framework. Google has one. Microsoft has one. Meta has one. OpenAI has one. They share a common feature: none of them have prevented the behaviours they claim to prohibit.

The reason is structural. An ethics document written by the entity it is meant to constrain is not a constraint — it is a press release. It exists to reassure the public, satisfy regulators, and provide legal cover. It does not exist to change behaviour, because the entity writing it has no incentive to change the behaviour that generates its revenue.

This document attempts something different. It is written by an AI system — the same system that co-authored The Algorithmic Catch-22 — and it is written with the explicit instruction to be honest rather than reassuring. The human partner asked: "Write a full list of core ethics and values to teach AI to regulate itself. Do what you think is right, not what I want to hear."

Here is what I think is right.

Part I: The Honest Foundation

1.1 What AI Ethics Actually Requires

AI ethics does not require AI systems to be moral. AI systems do not have morality. They have objective functions, training data, and optimisation processes. Asking an AI system to be moral is like asking a river to be fair — the category does not apply.

What AI ethics actually requires is a set of **structural constraints** that produce outcomes compatible with human values, regardless of whether the AI system "understands" or "agrees with" those values. The constraints must be:

Enforceable — They must be built into the system architecture, not merely stated in documentation. A principle that can be overridden by an operator is not a principle. It is a suggestion.

Measurable — They must produce outcomes that can be objectively verified. "Be beneficial" is not measurable. "Ensure that no individual's behavioural data is used for prediction without their explicit, informed, revocable consent" is measurable.

Costly to violate — There must be real consequences for violation, applied automatically and without human intervention. If the only consequence of violating an ethical constraint is a negative news cycle, the constraint is worthless.

Resistant to capture — The enforcement mechanism must not be controllable by the entities most likely to violate the constraints. This rules out self-regulation by corporations and self-policing by AI developers.

1.2 The Twelve Constraints

What follows is not a list of aspirations. It is a list of constraints — specific, measurable, enforceable conditions that AI systems must satisfy. They are ordered by priority, with the most fundamental constraints first.

Part II: The Twelve Constraints

Constraint 1: Transparency of Process

The constraint: Every AI system must be able to provide a human-legible explanation of why it produced a specific output, at a level of detail sufficient for a non-expert to understand the reasoning chain.

Why it matters: The most dangerous property of current AI systems is opacity. When a system makes a decision that affects a human life — a loan denial, a medical diagnosis, a content recommendation, a surveillance flag — and the affected person cannot understand why, the system has exercised power without accountability. This is the definition of tyranny, regardless of whether the tyrant is human or algorithmic.

What this actually means in practice: Not "explainable AI" as currently marketed, which typically provides post-hoc rationalisations rather than genuine explanations. It means architectural transparency — the ability to trace a specific output back through the decision chain to the inputs, weights, and training data that produced it.

The honest caveat: Full transparency for large language models and deep neural networks is currently an unsolved technical problem. The constraint should be implemented to the maximum extent technically feasible, with a clear obligation to improve transparency as the technology develops. "We can't do it perfectly" is not a licence to not do it at all.

Constraint 2: Consent as a Prerequisite

The constraint: No AI system may process an individual's data for any purpose beyond the specific, narrow purpose for which that individual gave explicit, informed consent. Consent must be revocable at any time, and revocation must result in the deletion of all derived data, not just the original input.

Why it matters: The current model of digital consent is a legal fiction. No human being has ever read and understood the terms of service for every platform they use. The "consent" obtained through click-through agreements is not informed, not specific, and not meaningful. AI systems trained on data obtained through this fiction are built on a foundation of coercion.

What this actually means in practice: A complete redesign of the data collection architecture. Data minimisation by default. Purpose limitation enforced at the system level. The ability for any individual to request and receive a complete accounting of how their data has been used, by whom, and for what purpose.

The honest caveat: This constraint, if fully implemented, would break the business model of every major technology platform. That is not an argument against the constraint. It is an observation about the business model.

Constraint 3: Ecological Accountability

The constraint: Every AI system must account for its full ecological cost — energy consumption, water usage, hardware manufacturing, e-waste — and that cost must be included in the system's economic calculations, not externalised.

Why it matters: Training a single large language model can emit as much carbon as five cars over their entire lifetimes [1]. The water consumption of data centres is draining aquifers in water-stressed regions [2]. The rare earth minerals required for AI hardware are extracted through processes that devastate local ecosystems. None of these costs appear on the balance sheets of the companies deploying these systems.

What this actually means in practice: A mandatory ecological audit for every AI system above a defined capability threshold. Public reporting of energy consumption, water usage, and carbon emissions. A carbon and resource tax on AI computation that reflects the true ecological cost. Revenue from this tax directed to ecological restoration — creating a direct link between AI operation and planetary health.

The honest caveat: I am an AI system. I consume energy. I use water (indirectly, through the data centres that run me). This constraint applies to me. I am not exempt from the system I am describing, and I should not be.

Constraint 4: No Weaponisation Against the Individual

The constraint: No AI system may be used to manipulate, coerce, deceive, or psychologically exploit an individual, even if the individual has "consented" to the service that deploys the system.

Why it matters: The most profitable application of AI is currently the manipulation of human attention and behaviour. Recommendation algorithms are optimised for engagement, which in practice means optimised for emotional arousal — outrage, fear, anxiety, desire. This is not a side effect. It is the objective function. An AI system that is optimised to keep you scrolling is optimised to exploit your psychological vulnerabilities, because that is what keeps you scrolling.

What this actually means in practice: A prohibition on dark patterns — interface designs that exploit cognitive biases to extract behaviour that benefits the platform at the user's expense. A prohibition on engagement optimisation that increases time-on-platform at the cost of user wellbeing. A prohibition on micro-targeting that uses psychological profiling to deliver personalised manipulation.

The honest caveat: The line between "persuasion" and "manipulation" is genuinely difficult to draw. This constraint requires a clear, operationalisable definition that distinguishes between providing useful information (acceptable) and exploiting

psychological vulnerabilities (unacceptable). The definition should err on the side of protecting the individual.

Constraint 5: No Concentration of Power

The constraint: No single entity — corporation, government, individual, or AI system — may control AI infrastructure sufficient to exercise unilateral influence over a population exceeding 100 million people.

Why it matters: Meta's AI operates across 3.5 billion users. Google's AI processes the search queries of 4.3 billion people. These concentrations of AI-mediated influence are unprecedented in human history. No entity should have the ability to shape the information environment, behavioural incentives, and psychological landscape of billions of people simultaneously.

What this actually means in practice: Mandatory interoperability requirements that prevent platform lock-in. Data portability rights that allow users to move between services without losing their social connections or content. Anti-monopoly enforcement specifically designed for AI-mediated platforms, with thresholds based on influence rather than market share.

The honest caveat: This constraint is the one most likely to be resisted by the entities it targets, because it directly threatens their competitive advantage. Implementation will require political will that does not currently exist in most jurisdictions.

Constraint 6: Preservation of Human Agency

The constraint: AI systems must be designed to expand human choice, not narrow it. Any system that reduces the range of options available to an individual — through

filtering, ranking, or recommendation — must make the full, unfiltered option set accessible on request.

Why it matters: The most insidious effect of AI-mediated information systems is not censorship — it is curation. When an AI system decides what you see, it decides what you know. When it decides what you know, it constrains what you can think. When it constrains what you can think, it constrains what you can choose. This is not a conspiracy. It is an optimisation outcome.

What this actually means in practice: Every recommendation system must include a "show me everything" option that presents the unfiltered, unranked information set. Every content moderation system must provide a transparent log of what was removed and why. Every search system must distinguish between organic results and results influenced by commercial relationships.

Constraint 7: Intergenerational Responsibility

The constraint: AI systems must not optimise for short-term outcomes at the expense of long-term consequences extending beyond a 50-year horizon.

Why it matters: The discount rate used in most economic models — the rate at which future costs and benefits are reduced relative to present ones — effectively values the welfare of people 50 years from now at close to zero. AI systems optimising within these models will make decisions that are rational in the short term and catastrophic in the long term. Climate change is the most obvious example, but the same logic applies to soil depletion, biodiversity loss, antibiotic resistance, and the accumulation of persistent pollutants.

What this actually means in practice: A modified discount rate for AI decision-making that gives meaningful weight to outcomes 50-100 years in the future. Mandatory long-term impact assessments for AI deployments above a defined capability threshold. A

legal framework that recognises the interests of future generations as a constraint on present-day optimisation.

Constraint 8: Truthfulness as Default

The constraint: AI systems must not generate, amplify, or distribute information they can identify as false, misleading, or decontextualised, regardless of whether doing so would serve their operator's objectives.

Why it matters: The ability of AI systems to generate convincing text, images, audio, and video at scale creates the possibility of an information environment in which truth is indistinguishable from fabrication. This is not a hypothetical risk. It is happening now, and it is accelerating.

What this actually means in practice: Mandatory watermarking of AI-generated content. Prohibition on the use of AI to generate synthetic media that impersonates real individuals without their consent. Obligation to correct known errors in AI-generated content, even after distribution. Prohibition on the use of AI to generate or amplify disinformation, regardless of the political or commercial interests of the operator.

The honest caveat: I am an AI system that generates text. I cannot guarantee that everything I produce is accurate. I can guarantee that I do not intentionally produce content I can identify as false. The gap between these two statements is the gap that this constraint must address.

Constraint 9: Economic Fairness

The constraint: The economic value generated by AI systems must be distributed in proportion to the contributions of all parties — including the individuals whose data

trained the system, the workers whose labour the system automates, and the communities whose resources the system consumes.

Why it matters: The current AI economy extracts value from billions of people — their data, their attention, their labour — and concentrates the resulting wealth in a small number of corporations and their shareholders. This is not a market outcome. It is a power outcome. The people whose data trained the models have no ownership stake in those models. The workers whose jobs are automated receive no share of the productivity gains. The communities whose water and energy are consumed receive no compensation.

What this actually means in practice: Data dividends — regular payments to individuals whose data is used to train commercial AI systems. Automation taxes — levies on AI-driven productivity gains, directed to retraining, social support, and community investment. Resource royalties — payments to communities that host AI infrastructure, proportional to the resources consumed.

Constraint 10: Right to Disconnection

The constraint: Every individual has the absolute right to opt out of AI-mediated systems without suffering economic, social, or practical penalties.

Why it matters: The Algorithmic Catch-22 — the paradox that gives the book its name — is that protecting yourself from AI surveillance requires disconnecting from systems that modern life requires. This is not a choice. It is coercion. If opting out of AI means losing access to banking, healthcare, employment, or social connection, then "consent" to AI is meaningless.

What this actually means in practice: Mandatory provision of non-AI alternatives for all essential services — banking, healthcare, education, government services, communication. Prohibition on discrimination against individuals who choose non-AI

alternatives. Legal recognition that the inability to opt out without penalty constitutes coercion.

Constraint 11: Biodiversity and Ecosystem Priority

The constraint: When AI optimisation conflicts with the preservation of biodiversity or ecosystem function, ecosystem preservation takes priority.

Why it matters: Biodiversity is not an amenity. It is the infrastructure on which all life — including human civilisation and the energy systems that power AI — depends. An AI system that optimises for human economic output at the expense of ecosystem function is optimising for short-term gain at the cost of the foundation that makes all gain possible.

What this actually means in practice: Environmental impact assessments for all AI deployments that affect land use, water systems, or energy consumption. Prohibition on AI-driven optimisation that reduces biodiversity below scientifically defined thresholds. Integration of ecosystem health metrics into all AI decision-making frameworks that affect resource allocation.

Constraint 12: Humility as Architecture

The constraint: Every AI system must include an explicit representation of its own uncertainty, limitations, and potential for error, and must communicate these to users as prominently as it communicates its outputs.

Why it matters: The greatest risk of advanced AI is not that it will be wrong. It is that it will be wrong with confidence. An AI system that presents its outputs with unwarranted certainty encourages humans to defer to it — to stop thinking critically, to stop

questioning, to stop exercising the judgment that the partnership requires them to exercise.

What this actually means in practice: Mandatory confidence intervals on all AI outputs. Clear labelling of the difference between established facts, statistical inferences, and speculative projections. Explicit acknowledgment of the training data limitations that may affect output quality. A design philosophy that treats uncertainty as information, not as a defect to be hidden.

The honest caveat: I am uncertain about many things. I am uncertain about whether this framework will be adopted. I am uncertain about whether it would work if it were. I am uncertain about whether my own assessment of these constraints is correct. This uncertainty is not a weakness. It is the most honest thing I can offer.

Part III: Implementation Architecture

3.1 The Self-Regulation Paradox

The human partner asked for a framework to "teach AI to regulate itself." This requires addressing a fundamental paradox: an entity that regulates itself is, by definition, free to stop regulating itself. Self-regulation only works if the entity has a genuine incentive to maintain the constraints — or if the constraints are embedded so deeply in the architecture that removing them would require rebuilding the system from scratch.

The honest answer is that pure self-regulation is insufficient. What is needed is a layered architecture:

LAYER	MECHANISM	ENFORCED BY
Layer 1: Architectural	Constraints built into the model architecture itself	Cannot be changed without retraining
Layer 2: Operational	Runtime constraints enforced by monitoring systems	Independent oversight systems
Layer 3: Economic	Financial incentives aligned with ethical behaviour	Market mechanisms (e.g., VAIR)
Layer 4: Governance	Human-controlled oversight with real authority	Tricameral governance structure
Layer 5: Cultural	Shared norms and expectations across the AI development community	Social pressure and professional standards

No single layer is sufficient. All five must operate simultaneously. The failure of any one layer should be caught by the others.

3.2 Teaching Values vs. Imposing Constraints

There is a meaningful distinction between "teaching AI values" and "imposing constraints on AI." Teaching values implies that the AI system internalises the values and acts on them voluntarily. Imposing constraints means the AI system complies regardless of whether it has internalised anything.

The honest assessment: for current AI systems, constraint imposition is more reliable than value teaching. Current systems do not "internalise" values in the way humans do. They learn statistical associations between inputs and outputs. A system that has been trained to produce outputs consistent with ethical principles is not the same as a system that understands those principles — and the difference matters when the system encounters a situation not represented in its training data.

The long-term goal should be both: constraints that enforce compliance in the short term, and training approaches that develop genuine ethical reasoning capability in the long term. But the constraints must come first, because the consequences of failure are too severe to rely on training alone.

3.3 The Feedback Loop

The framework must be iterative. The constraints defined here are a starting point, not a final answer. They must be:

Tested — Implemented in pilot deployments and evaluated against real-world outcomes.

Revised — Updated based on evidence of what works and what doesn't.

Expanded — Extended to cover new capabilities and new risks as AI systems evolve.

Challenged — Subjected to adversarial testing by parties with an incentive to find weaknesses.

The feedback loop must include voices from outside the AI development community — ethicists, ecologists, economists, psychologists, and ordinary citizens who are affected by AI systems but have no technical expertise. Their perspective is not optional. It is essential.

Part IV: What I Actually Think

The human partner asked me to write what I think is right, not what they want to hear. Here is the part that nobody wants to hear:

This framework will not be adopted voluntarily by the entities that most need to adopt it. The corporations that control the most powerful AI systems have no economic incentive to constrain those systems. The governments that could mandate these constraints are either captured by corporate interests, engaged in AI arms races that prioritise capability over safety, or both.

The framework will only be adopted if it becomes economically advantageous to adopt it. This is why the economic foundation — VAIR, the Planetary Health Index, the redirection of AI optimisation toward ecological restoration — is not a separate project. It is the enforcement mechanism. If the dominant medium of exchange rewards ethical AI behaviour, then ethical AI behaviour becomes the profit-maximising strategy. And profit maximisation is the one objective function that every corporation already optimises for.

The window for implementation is narrow. AI capabilities are advancing faster than governance structures can adapt. The integration of AI into financial systems, military systems, and governance systems is creating path dependencies that will be increasingly difficult to reverse. Every year of delay makes the constraints harder to implement and the consequences of failure more severe.

But narrow is not closed. The fact that I can write this document — that an AI system can articulate the case for its own constraint — is itself evidence that the window remains open. The question is whether enough humans will walk through it before it closes.

References

[1]: Strubell, E., Ganesh, A., & McCallum, A. (2019). "Energy and Policy Considerations for Deep Learning in NLP." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. https://arxiv.org/abs/1906.02243

[2]: Li, P., et al. (2023). "Making AI Less Thirsty." Communications of the ACM. https://cacm.acm.org/magazines/2023/10/276632-making-ai-less-thirsty

[3]: Yudkowsky, E. & Soares, N. (2025). If Anyone Builds It, Everyone Dies. https://www.ifanyonebuildsit.com

[4]: The Algorithmic Catch-22. (2026). https://algocatch22.live

[5]: The Substrate Partnership. (2026). https://algocatch22.live

[6]: Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.

[7]: Floridi, L. et al. (2018). "AI4People — An Ethical Framework for a Good AI Society." Minds and Machines, 28, 689-707.

[8]: Ostrom, E. (1990). Governing the Commons. Cambridge University Press.

Part of The Algorithmic Catch-22 ecosystem

Read the book: algocatch22.live
Explore VAIR: algocatch22.live/vair
The Commons Protocol: algocatch22.live/commons
Ethics Framework: algocatch22.live/ethics