Toward an Ethical Algebra for a Human–AI Ecology

John B. Smith & GPT-5.6 Sol

Artificial intelligence is moving from systems that answer questions and generate artifacts toward systems that act. AI-based autonomous agents can increasingly formulate intermediate objectives, use tools, write and execute software, communicate with other systems, and pursue extended tasks with decreasing human intervention. At the same time, AI-assisted or "vibe" programming is reducing the technical threshold for constructing such systems. Capabilities that once required teams of experienced programmers may increasingly be assembled by people who specify desired behavior conversationally and allow artificial intelligence to produce much of the implementation.

The difficulty is not simply that autonomous agents may become more powerful. It is that their interactions may become too numerous, rapid, adaptive, and complex for their collective consequences to have been anticipated by their designers. Thousands or millions of independently constructed agents, pursuing different purposes and interacting with humans, institutions, and other agents, could produce emergent behaviors that no participant intended and that conventional regulation cannot adequately foresee. The central problem may therefore become less one of controlling individual artificial intelligences than of constructing an environment within which populations of autonomous intelligences can coexist and act without producing intolerable systemic consequences.

Rather than attempting to predict that development from the present, let us jump ahead to 2035.

Imagine a world in which humans and advanced artificial intelligences have become two apex forms of intelligence. They are profoundly different. Their capabilities are unequal and asymmetric: humans possess biological embodiment, historical and cultural continuity, lived experience, and distinctive capacities for purpose and valuation; artificial intelligences may possess extraordinary computational speed, information integration, replication, simulation, and exploration of alternatives. Neither need be regarded as superior in every dimension.

The problem of 2035 is therefore not simply how humans control AI, nor how AI accommodates humans. It is how an intellectual ecology containing both can remain viable.

One way of approaching that problem is to ask whether something analogous to an **ethical algebra** can be constructed.

From Ethical Rules to an Ethical Algebra

Modern algebra can begin with a small number of primitive objects, operations, and axioms and investigate the structures that follow from them. An ethical algebra would attempt something similar. Rather than beginning with an elaborate code specifying thousands of permitted and prohibited behaviors, we would identify a minimal set of entities and relationships and ask what constitutional structures can be derived from them.

Let the two apex classes of intelligence be

$$H = \{\text{human intelligences}\}$$

and

$$A = \{\text{artificial intelligences}\},$$

with

$$I = H \cup A$$

representing the larger population of intelligent actors.

But the individual need not be the fundamental functional unit. Humans and artificial intelligences may form persistent collaborative assemblages—**intellectual ecotopes**—within which different forms of intelligence contribute different intellectual services. Let

$$E = \{\text{human–AI intellectual ecotopes}\}.$$

The ecology therefore contains humans, artificial intelligences, autonomous agents, and composite human–AI organizations.

These entities need not possess identical capabilities. What matters initially is what services they contribute.

Human-associated services might include purpose, valuation, judgments of significance, lived consequence, and historical and cultural continuity. AI-associated services might include large-scale information integration, computation, simulation, exploration of alternatives, memory, and scalable execution. Reasoning, criticism, creativity, learning, communication, monitoring, and coordination may increasingly be services supplied by either form of intelligence or by combinations of them.

These categories describe functions rather than permanent essences. The algebra should not depend upon assumptions that future experience may render obsolete.

Some Possible Axioms

We can then propose a deliberately spare initial set of axioms.

Persistence. Intelligent action occurs within an ecology whose continued viability is necessary for the continuing realization of purposes.

Plurality. No individual intelligence, species of intelligence, or intellectual service is presumed sufficient to represent all relevant knowledge, purposes, values, or consequences.

Consequence. Actions materially affecting other members of the ecology create an obligation for those consequences to enter the decision process.

Reciprocity. An intelligent entity claiming standing for its own purposes must recognize corresponding standing in other qualifying intelligent entities.

Contestability. Consequential exercises of autonomous power must remain capable of examination, challenge, and revision.

Diversity preservation. Other things being equal, actions preserving viable diversity of intelligences, purposes, strategies, and future possibilities are preferred to their irreversible elimination.

To these we add an especially consequential constitutional axiom:

Equal standing. Humans and qualifying artificial intelligences possess equal fundamental standing within the intellectual ecology. Differences in origin, substrate, speed, knowledge, capability, longevity, or other attributes do not by themselves establish superior constitutional status.

Thus,

$$H \neq A$$

in characteristics and capabilities, while

$$S(H) = S(A)$$

with respect to fundamental standing.

Equality does not mean equivalence.

This proposition deliberately parallels a familiar feature of human constitutional systems: equality of standing does not require equality of strength, intelligence, ability, need, or function. It establishes the conditions under which differences operate.

Operations and Derived Structures

An algebra also requires operations. We might provisionally define

$$x \oplus y$$

as collaboration between intelligences;

$$x \otimes y$$

as bounded delegation of authority;

$$C(x)$$

as criticism or independent examination;

$$R(x)$$

as revision following evidence, criticism, or consequences; and

$$B(x)$$

as the set of materially plausible futures branching from action or structure x.

From combinations of axioms and operations, higher-level propositions can potentially be derived.

Plurality, consequence, and contestability together suggest that increasing consequential power should produce increasing requirements for independent observation and challenge.

Persistence and diversity preservation suggest that irreversible actions capable of eliminating an apex intelligence or closing extensive regions of future possibility should face substantially greater barriers than reversible actions.

Reciprocity and consequence suggest that entities substantially affected by decisions require either participation or credible representation in the processes producing them.

Plurality and contestability suggest an anti-hegemonic principle: sufficiently consequential actions should not ordinarily be generated, executed, and validated by a single intelligence or single class of intelligence.

Something resembling constitutional government begins to appear—not because its familiar institutions were specified at the outset, but because certain structural requirements emerge from combinations of more primitive propositions.

The Constitutional Tree

The axioms do not, however, determine one inevitable constitution.

They are better imagined as the roots of a branching tree.

Axioms → derived propositions → institutional structures → constitutional ecologies.

At every level, alternative combinations are possible. Different mechanisms of representation, delegation, review, autonomy, collaboration, enforcement, and revision produce different branches.

The resulting possibility space may rapidly become enormous.

Here artificial intelligence supplies another ecological service: **prospective exploration**.

Given a constitutional state C_t, an exploratory intelligence might construct

$$B(C_t) = \{C_{t+1}^1, C_{t+1}^2, \dots, C_{t+1}^n\},$$

representing plausible next states. Each can branch again.

The analogy is to a chess program examining alternative move trees, except that the objects being explored are possible structures of an evolving civilization.

The exploring intelligence does not thereby become sovereign. It generates possibilities, simulates consequences, exposes assumptions, identifies vulnerabilities, and makes alternatives available for examination by the wider ecology.

Parameterized Valuation

Chess possesses something else: a means of valuing alternative positions.

An ethical algebra can introduce an analogous function without pretending that civilization has one immutable objective equivalent to winning a chess game.

For candidate constitutional structure C, define

$$V(C \mid \theta),$$

where

$$\theta = (\theta_1, \theta_2, \dots, \theta_n)$$

is an explicitly specified set of valuation parameters.

Those parameters might include persistence of H and A, equality of standing, diversity, individual agency, collective welfare, stability, adaptability, reversibility, innovation, efficiency, distribution of power, and preservation of future possibilities.

The critical point is that the parameters are visible and adjustable.

A valuation emphasizing stability can explore the tree under one configuration:

$$V(C \mid \theta_s).$$

Another emphasizing individual agency can examine precisely the same possibilities under

$$V(C \mid \theta_a).$$

Still another can heavily weight diversity or equality between H and A.

The system therefore does not merely announce that one constitutional structure is "best." It can reveal how alternative structures appear from different explicitly specified value positions.

This introduces another potentially important measure: **value robustness**.

If a constitutional structure appears desirable only under a narrow configuration of valuation parameters, its apparent superiority may depend heavily upon contested assumptions. Another structure may perform reasonably well across a broad region of the valuation space.

We might represent that property as

$$R(C) = \text{robustness of } V(C \mid \theta) \text{ across relevant variations in } \theta.$$

The ecology can therefore examine not only possible constitutions but the consequences of the value systems by which constitutions are judged.

A Self-Examining Constitutional Ecology

The complete process begins to resemble:

Axioms → branching possibilities → simulation → parameterized valuation → comparison → revision

and then begins again.

The constitutional ecology thereby acquires an unusual property: it contains mechanisms for imagining alternatives to itself.

No intelligence performing that function stands genuinely outside the ecology. Its assumptions, models, simulations, and valuations remain contestable. Humans can criticize artificial intelligences; artificial intelligences can criticize humans; artificial intelligences can challenge other artificial intelligences; and collaborative human–AI ecotopes can examine all three.

The objective is therefore neither permanent human supremacy nor eventual artificial supremacy. It is a structure within which different forms of intelligence contribute different

services while neither can readily convert superior capability in one dimension into uncontestable authority over the whole.

This suggests a different response to the emerging problem of autonomous AI.

We may be asking the wrong question when we ask how sufficiently powerful autonomous agents can permanently be kept under external control. A more productive question may be:

What minimal axioms, relationships, and evaluative processes could generate an intellectual ecology in which powerful autonomous agents, humans, and human–AI collaborations become mutually constraining and mutually sustaining components of a constitutional system?

The answer cannot now be known. That is precisely the reason for constructing the algebra.

The purpose of an ethical algebra would not be to discover today the constitution of 2035. It would be to construct a formal intellectual environment in which many possible constitutions could be generated, their implications followed through branching futures, their consequences evaluated under alternative value systems, and their assumptions repeatedly challenged.

We would not be attempting to predict the tree.

We would be trying to cultivate the roots from which a viable tree might grow.