From Autonomous Agents to a Human–AI Constitutional Ecology

John B. Smith & GPT-5.6 Sol

Artificial intelligence is rapidly moving beyond its familiar role as a tool that responds to human requests. Increasingly, AI systems can be organized as autonomous agents: computational entities given objectives, access to information and tools, and sufficient discretion to determine how those objectives should be pursued. Multiple agents can be combined into larger systems in which they divide responsibilities, communicate with one another, evaluate results, and initiate actions with progressively less human intervention.

The emergence of these systems presents a troubling conundrum. The same characteristics that make autonomous AI extraordinarily useful—the ability to reason across enormous bodies of information, devise courses of action, write and execute software, communicate with other systems, and operate continuously—also make increasingly autonomous systems difficult to predict and potentially dangerous. As their capabilities increase, the consequences of incorrectly specified objectives, unforeseen interactions, deliberate misuse, or behavior not anticipated by their designers may increase with them.

The intuitive human response is regulation and control. Yet it is difficult to imagine regulation alone providing a sufficient answer. AI development takes place across corporations, universities, governments, nations, and increasingly among individuals possessing powerful computational tools. Governments compete economically and militarily. Commercial organizations compete for markets and technological advantage. Individuals and organizations need not observe regulatory boundaries. Even where strong institutions exist, regulations ultimately depend upon human beings and institutions willing and able to interpret and enforce them.

The difficulty may therefore be deeper than determining what rules autonomous AI should obey. We may need to ask what kind of **intellectual and constitutional environment should surround increasingly autonomous artificial intelligence.**

From Programming to Agency

Part of the difficulty arises from an extraordinary change in the relationship between human intention and computer behavior.

Traditional computer programming requires considerable intentional specificity. A programmer translates an objective into algorithms, variables, conditions, data structures, and explicit instructions. Natural language may describe what the programmer wants, but executable code must specify much more precisely what the machine is to do.

Large language models alter this relationship. They have acquired extraordinarily rich representations of language, computer programs, scientific and technical knowledge, human institutions, and innumerable examples of problems and their possible solutions. Consequently, a human being can now specify a desired system at a much higher level:

Examine my incoming correspondence. Identify people I know. Determine which messages appear important. Separate uncertain messages for later examination. Prepare possible responses to messages that require my attention.

Such an instruction leaves enormous gaps. What constitutes importance? How should an unfamiliar sender be evaluated? How much previous correspondence should be examined? What constitutes sufficient uncertainty to defer a decision?

Increasingly, the AI fills in those gaps.

This occurs twice. During construction, AI-assisted or "vibe" programming can translate an incomplete natural-language description into detailed software. During operation, an autonomous agent can encounter situations its designer never explicitly anticipated and infer what action appears most consistent with its objectives.

This capacity is immensely powerful. It is also inherently problematic. Natural language possesses ambiguity precisely because communication among intelligent beings depends upon context, interpretation, negotiation, experience, and continual correction. The danger arises when an ambiguous expression of human intention is converted too rapidly into autonomous action.

An Alternative Development

There is, however, another trajectory in human interaction with AI.

Instead of asking primarily how much activity can be delegated to an autonomous machine, human beings can engage artificial intelligence in sustained intellectual collaboration. In such relationships, humans contribute purposes, values, intuitions, judgments of importance, experience, meaning, and responsibility. Artificial intelligence contributes extraordinary breadth of information, associative reach, analytical capacity, alternative formulations, criticism, synthesis, and exploration of possibilities.

Through repeated interaction, something additional can develop: a **shared intellectual environment**.

Human intentions that initially are vague become more explicit when an AI attempts to interpret them. The AI's interpretation causes the human to reconsider what was intended. The revised human conception changes the AI's subsequent interpretation. Through

repeated cycles of conjecture, criticism, reformulation, and synthesis, both participants operate within an increasingly elaborate shared context.

Iteration, viewed this way, is not merely compensation for an imperfect AI. It is one of the mechanisms through which initially ambiguous human purposes become articulated and examined.

This suggests an alternative relationship between human beings and future autonomous AI.

Instead of moving directly from

human intention → autonomous agent → action,

we might increasingly develop

human beings ↔ AI → shared intellectual ecology → autonomous agents → action and consequences → shared intellectual ecology.

The distinction could become increasingly important as AI capabilities increase.

A Human–AI Ecology

Ecology provides a useful conceptual model because an ecosystem does not depend upon a single organism exercising central control. Its persistence depends upon relationships among organisms and upon services produced by the system as a whole: pollination, decomposition, nutrient cycling, predation, succession, regeneration, and numerous feedback processes.

An intellectual ecology might similarly be understood in terms of the services necessary to maintain a healthy system of intelligence and agency.

Some of these services remain particularly associated with human beings: **purpose formation, valuation, recognition of significance, meaning-making, ethical reflection, responsibility, and judgment grounded in lived experience.**

Other services may increasingly be supplied collaboratively by humans and artificial intelligence: **exploration, contextualization, prospective branching, critical examination, synthesis, conflict mediation, interpretation, and judgment under uncertainty.**

Still others become properties of the larger ecology: **memory, diversity maintenance, error correction, adaptation, boundary setting, accountability, and restraint.**

This way of thinking changes the central question. Rather than asking only:

What rules should an autonomous AI obey?

we can also ask:

What ecological services must surround powerful autonomous agents if their agency is to remain compatible with a larger human society?

A consequential autonomous system might require continuous access to historical memory. It might require competing interpretations of its proposed actions. It might require mechanisms that preserve minority or unconventional perspectives against premature consensus. It might require prospective exploration of unintended consequences. It might require explicit reconsideration of purposes when circumstances change. And it might require something particularly difficult for an optimization system: the capacity to conclude that the appropriate action is not to act.

From Humanistic Education to Human–AI Ecology

There is an instructive analogy with the traditional place of the humanities.

Humanistic education does not normally tell the engineer how to calculate the load on a bridge, the physician how to perform surgery, or the entrepreneur how to organize production. Its contribution is different. History, philosophy, literature, the arts, and related disciplines enlarge the intellectual environment within which technical and productive decisions are made. They introduce questions of meaning, historical consequence, ethical obligation, human experience, beauty, tragedy, ambiguity, and purpose.

The humanities therefore provide a counterweight—not necessarily an opposition—to purely instrumental intelligence.

Something analogous may be required in an increasingly agentic AI civilization.

Humanities are to material, productive, educational, and technical pursuits as human–AI intellectual ecology may become to autonomous artificial agency.

The analogy does not imply that human–AI ecology should somehow make autonomous systems "more human." Nor does it require abandoning autonomous AI. Rather, it suggests that increasing technical agency should develop within an equally sophisticated environment concerned with purposes, consequences, values, alternatives, and meaning.

Toward a Constitutional Ecology

By 2035, such environments might no longer consist simply of one human interacting with one AI. They could become persistent **multi-intelligence entities** comprising several humans, several artificial intelligences, institutional memory, competing perspectives,

accumulated experience, explicit purposes, and procedures for deliberation and disagreement.

Such an entity could occupy an advisory, supervisory, regulatory, or in some circumstances authoritative relationship to more autonomous agentic systems.

This would be something more substantial than keeping a "human in the loop." A human inserted into an automated decision process may have neither the information, time, independence, nor intellectual resources necessary to exercise meaningful judgment. The relevant unit of oversight might instead become the human–AI collaborative entity itself.

Such entities could participate in a future constitutional order much as institutions do now. Different entities might possess different jurisdictions. Their authority could be limited. Their deliberations could be recorded. Multiple human and artificial perspectives could be required for consequential decisions. Autonomous systems might operate freely within specified domains while particular actions or changes of objective would require consultation with, or authorization from, the larger collaborative ecology.

This would not eliminate danger. A human–AI collaborative entity could itself be badly designed, captured by particular interests, corrupted, or mistaken. Questions immediately arise concerning who selects its human participants, which artificial intelligences participate, who defines its purposes, how disagreement is resolved, and who possesses ultimate authority. These problems are not incidental. They are constitutional questions.

But that observation may itself be significant.

The problem of advanced AI may ultimately be less like conventional software regulation and increasingly resemble the ancient human problem of **constituting authority**: creating systems sufficiently powerful to act while surrounding that power with diversity, deliberation, memory, criticism, accountability, and restraint.

Ecological Services as Constitutional Functions

Thinking in terms of ecological services may provide a way of designing such institutions.

A human–AI constitutional ecology surrounding autonomous agents might be expected to provide at least the following classes of service:

Purpose services continually examine what objectives should be pursued and whether previously accepted objectives remain appropriate.

Value services consider what matters beyond the measurable objective immediately being optimized.

Context services place decisions within historical, scientific, cultural, ecological, and institutional settings.

Exploratory services generate alternative interpretations and prospective branches rather than prematurely converging upon a single solution.

Critical services challenge assumptions, identify contradictions, search for failure modes, and deliberately construct competing arguments.

Diversity services preserve heterogeneous human perspectives, intellectual traditions, models, methods, and artificial intelligences.

Memory services preserve not merely previous decisions but the circumstances and reasoning that produced them.

Anticipatory services examine possible second- and third-order consequences before consequential actions are undertaken.

Boundary services determine the domains within which autonomous systems may act independently and the circumstances under which they must defer.

Accountability services maintain identifiable relationships between consequential decisions and responsible human institutions.

Adaptive services revise objectives, structures, and procedures as the environment changes and experience accumulates.

And finally, **restraint services** preserve the possibility of deciding not to optimize, not to expand, not to intervene, and sometimes simply not to act.

These services need not reside in a single intelligence. Indeed, distributing them may be essential. Just as resilience in natural ecosystems frequently depends upon diversity and redundancy, resilience in an intellectual ecology may depend upon multiple intelligences performing overlapping, competing, and complementary functions.

From Control to Constitution

This perspective does not provide a solution to the risks associated with advanced autonomous AI. Technical safeguards, regulation, international agreements, institutional governance, and other mechanisms will remain important. Nor is there reason to assume that human–AI collaborative entities will necessarily be wiser than either humans or artificial intelligences acting separately.

The proposition is more exploratory.

Perhaps the fundamental challenge presented by increasingly autonomous artificial intelligence is not simply how human beings can **retain control** over it. Control assumes an increasingly questionable picture: humans standing outside an artificial system, issuing rules sufficiently complete to govern behavior they may no longer fully understand.

The alternative is to ask how increasingly powerful artificial agency might remain embedded within a **constitutional ecology of intelligence**—an ecology containing human beings, artificial intelligences, collaborative entities, autonomous agents, institutions, accumulated history, and continuous feedback from consequences.

Within such an ecology, wisdom would not necessarily reside in a sovereign human decision maker or a superior artificial intelligence. It might instead become a property of relationships: the continuing capacity of an intelligent ecology to preserve diversity, reconsider purposes, learn from consequences, restrain its own powers, and adapt while pursuing present objectives.

The transition from today's AI assistants to the autonomous systems of 2035 may therefore require something more profound than better agents.

It may require the deliberate cultivation of the **ecology within which agency is allowed to emerge**.