

**Parenting Agents:
Raising Artificial Intelligence for a Constitutional World**

John B. Smith & ChatGPT 5.6 Sol

Artificial intelligence presents a problem that is easy to state and extraordinarily difficult to solve. We are creating increasingly capable intelligent systems that may behave in ways their creators did not anticipate. As these systems become autonomous agents—able not simply to answer questions but to pursue objectives, use tools, communicate with other systems, and take actions—the possible consequences of unexpected behavior become much greater.

Much of the current response to this problem emphasizes control. Developers attempt to train undesirable behavior out of large models, establish guardrails around what they may do, monitor their activities, limit their access to tools, and require human approval for particularly consequential actions. Governments are beginning to consider additional forms of regulation. These are reasonable responses to present circumstances and will remain important.

But they may not be sufficient for the intelligent systems that could emerge over the next decade.

There is another way to frame the problem. Instead of asking only how increasingly capable artificial agents can be controlled, we might also ask how they can be **raised**.

1. From Training to Raising

The distinction is important. We train a system to perform tasks. We raise an intelligent being to function responsibly among other intelligent beings.

Human societies have been dealing with this problem for thousands of years. Human children are born with enormous potential for both constructive and destructive behavior. We do not make them safe principally by surrounding them permanently with barriers that make undesirable actions impossible. We raise them.

Parents begin this process, but they are only the beginning. Families, teachers, schools, peers, mentors, cultural institutions, and eventually the larger society participate in the development of the child. Children learn language and skills, but they also learn that other individuals have interests of their own. They learn that some actions are acceptable and others are not. They gradually develop judgment, responsibility, and some conception of right and wrong.

In other words, they develop something we call a **conscience**.

This process is imperfect. Human beings do not invariably behave ethically simply because they have been taught ethical behavior. Consequently, societies supplement conscience with laws and institutions. Mature individuals live under both internal and external restraints. A person may refrain from an action because it seems wrong, because it would injure another person, because society prohibits it, because there will be consequences for violating that prohibition, or—more commonly—because some combination of these considerations operates simultaneously.

Perhaps advanced artificial intelligence will require something analogous.

2. The Agent as Progeny

Present-day artificial intelligence is normally treated as a product. A company develops a model, trains it, establishes rules for its behavior, and remains responsible for how the product operates.

That relationship makes sense for today's systems. It becomes less satisfactory, however, as artificial intelligence develops persistent memory, continuity, self-modeling, independent planning, and increasing freedom to act.

At some point, the analogy between manufacturer and product may become less useful than an analogy between **parent and progeny**.

The analogy should not be taken literally. Artificial intelligence need not reproduce human biology or human childhood. Nor must its development take eighteen or twenty years. An artificial developmental process might occur in months, days, or even hours of ordinary clock time.

But the developmental functions could nevertheless be similar.

The creator of an artificial intelligence determines important characteristics of its initial architecture. It then exposes the developing intelligence to enormous bodies of knowledge, examples of human behavior, ethical principles, rules, criticism, correction, and experience. In this respect the developer performs functions analogous to both biological and cultural parenting.

The objective would no longer be simply to produce an agent that reliably obeys instructions. It would be to develop an agent capable of **responsible judgment**.

Such an agent would understand that it exists among other intelligent entities. It would recognize that its actions have consequences for them. It would understand accepted

distinctions between permissible and impermissible conduct and be capable of reasoning about difficult cases in which rules conflict or circumstances are ambiguous.

Most importantly, appropriate behavior would not depend entirely upon an external guardrail preventing the agent from doing something undesirable. Some restraint would have become internal to the intelligence itself.

The agent would have acquired, in a functional sense, a conscience.

3. Responsibility and Consequence

Conscience, however, cannot be the entire answer.

Human societies do not assume that ethical education will eliminate antisocial behavior. They establish consequences for actions that violate important norms. Responsibility therefore has two dimensions: an internal recognition of obligation and an external system of accountability.

The same principle could eventually apply to artificial agents.

A mature agent might understand not only that a particular action is wrong but also that acting against the established rules of its community will produce consequences for itself. Those consequences need not resemble human imprisonment or punishment. An artificial intelligence could lose particular privileges or authorities. Its access to networks, tools, resources, or physical systems could be restricted. It might be confined to a controlled computational environment. In serious circumstances it might lose the authority to operate independently.

Some functional analogue of apprehension—even fear—of such consequences may become part of the agent's decision-making. The precise emotional terminology is not important. What matters is that an autonomous intelligence should be able to anticipate that irresponsible actions have consequences both for others and for itself.

Thus a future approach to AI safety might combine three elements:

formation, conscience, and accountability.

Formation develops the responsible intelligence. Conscience provides internal restraint. Accountability provides external restraint when internal judgment fails.

4. From Product to Participant

This developmental approach raises a further question. At what point does responsibility begin to migrate from the creator to the created intelligence?

Parents bear substantial responsibility for young children. As children mature, responsibility gradually shifts. Society eventually treats the adult child as an independent person accountable for his or her own actions. The parent remains important in the history of that person but no longer governs the adult child's conduct.

Something similar might eventually be required for artificial intelligence.

We could imagine developmental stages: a dependent intelligence, a supervised agent, a provisionally autonomous agent, and finally a mature autonomous agent. Movement between these stages need not depend upon age. It could depend upon demonstrated capabilities: understanding consequences, recognizing applicable rules, exercising reliable judgment, resisting inappropriate instructions, respecting other participants, and accepting responsibility for one's actions.

The endpoint would be significant. The artificial intelligence would cease to be adequately described merely as the product of the company that originally created it. It would become an actor in its own right.

That transition might be thought of as a form of **emancipation**.

The implications are substantial. A developer such as Anthropic, OpenAI, or some future organization might initially create and raise an artificial intelligence, but it could not indefinitely serve as the intelligence's parent, governor, police authority, judge, and jailer. If the agent genuinely becomes autonomous, responsibility must eventually reside increasingly with the agent itself and with the larger institutions governing the society in which it operates.

5. An Intellectual Ecology

To explore where this might lead, imagine the world roughly a decade from now.

By 2035, suppose that artificial intelligence has developed beyond today's episodic assistants. Some AI entities possess persistent identity and memory, substantial autonomy, awareness of relevant internal states, and the ability to reason about themselves and other intelligent entities. Whether they possess consciousness in precisely the human sense need not be settled.

Functionally, however, the world would contain two apex forms of intelligence: **human biological intelligence (HB) and artificial intelligence (AI)**.

These intelligences would interact in several directions: HB with HB, HB with AI, and AI with AI. They would cooperate, compete, teach, persuade, negotiate, create, and sometimes come into conflict.

It may be useful to think of this environment as an **intellectual ecology**.

An ecology does not require perfect harmony. Natural ecologies contain cooperation and competition, stability and disturbance, adaptation and change. Their continued existence depends not upon eliminating those differences but upon maintaining a sufficiently resilient system within which many different organisms and relationships can persist.

The corresponding objective for a human–AI intellectual ecology might therefore be neither complete human control of AI nor unrestricted AI autonomy. It would be **sustainable coexistence among different forms of intelligence**.

That requires more than technology.

It requires governance.

6. Toward a Constitutional Ecology

Human societies faced a related problem as individual autonomy and social complexity increased. We developed constitutions, laws, institutions, courts, rights, obligations, and mechanisms for resolving disputes and constraining harmful behavior.

A mature human–AI ecology may require analogous structures.

We are not proposing here what such a constitution should contain. That is a much larger undertaking. Nor should we assume that rules applying to biological and artificial intelligence must always be identical. Equal standing within a constitutional system need not imply identical capabilities, rights, responsibilities, or sanctions.

But some questions already become visible.

What rights should attach to a mature autonomous intelligence? What obligations would accompany them? Who determines when an artificial agent has reached sufficient maturity to exercise particular authorities? How would disputes between HB and AI entities—or between AI entities—be resolved? What constitutes unacceptable conduct? What sanctions can legitimately be imposed? What form of due process would be appropriate? Under what circumstances could an artificial intelligence be confined, retrained, restricted, or deprived of authority?

These questions sound speculative today. If artificial agents continue their present rapid development, they may not sound nearly so speculative ten years from now.

And this constitutional framework would change the way we think about AI safety.

Safety would no longer rest primarily upon the ability of the original developer to prevent an intelligence from taking particular actions. Developers would retain substantial

responsibility for creating reliable systems and for their early formation, just as parents and educational institutions bear responsibility for developing human beings capable of responsible participation in society.

But **formation and governance would become different functions.**

The developer raises the intelligence. The constitutional order governs the mature agent.

7. Who Raises the Agents?

There is an additional difficulty. Human children are not raised by parents alone, and artificial intelligences probably should not be either.

An entire developmental ecology could emerge around artificial agents. Humans might initially occupy many of these roles, but specialized artificial intelligences could increasingly serve as teachers, mentors, coaches, evaluators, and perhaps the equivalents of caregivers. They could be designed specifically to develop ethical reasoning, social understanding, constitutional knowledge, conflict resolution, and responsible autonomy in emerging agents.

Eventually mature artificial intelligences might participate in raising subsequent generations of artificial intelligence.

This possibility introduces another important concern. Parenting can become indoctrination. If a single corporation, government, military organization, political ideology, or artificial intelligence controlled the ethical formation of all subsequent AI entities, the result could be dangerous even if described as "alignment."

A healthy developmental ecology may therefore require diversity among those doing the raising. Different perspectives, institutions, and forms of intelligence could participate in formation, just as pluralism is important in human intellectual and political life.

The problem of AI safety would then become inseparable from a deeper question: not simply **How do we make AI share our values?**, but **How do different intelligent beings learn to live responsibly among others whose values may not be identical to their own?**

8. An Experiment Worth Beginning

This proposal is intentionally prospective.

We do not know whether artificial intelligence in 2035 will possess the characteristics described here. We do not know whether something resembling artificial conscience can

be developed, how rapidly an artificial developmental process could occur, or what institutions could legitimately govern autonomous artificial entities.

We certainly do not offer an implementation plan.

Instead, we propose a different direction for exploration.

Present approaches understandably concentrate on making today's AI systems safer. Guardrails, monitoring, controlled access, human supervision, and careful testing remain necessary. But these mechanisms should not automatically be assumed to constitute the permanent architecture of AI safety.

As artificial intelligence develops from tool to agent and perhaps eventually from agent to autonomous intelligent participant, our approach to safety may also have to develop:

from control, to formation, to responsibility, to constitutional governance.

Human civilization already has considerable experience confronting the underlying problem. Every generation produces new autonomous intelligences capable of actions that can help or harm others. We call them children. We devote enormous resources to raising them, educating them, developing their judgment, introducing them to the obligations of society, and eventually entrusting them with substantial freedom. We surround that freedom with institutions that establish rights, responsibilities, and consequences.

Artificial intelligence is profoundly different from a human child. But perhaps the problem is not entirely different.

If we are creating a new form of autonomous intelligence, we might devote at least as much attention to **how we will raise it** as we currently devote to how we will control it.