

This is a sidebar to the longer, temporarily on-going project concerned with vibe writing. If we can, let's give this extension a label of "parentingAgents," perhaps to be relabeled later.

I want to develop this into a artifact – a document – to be collaboratively written, as we now commonly do. But, in this case I want it to be multistage with us first enumerating major ideas/concept, then identifying a focus and intent, and then an outline of the major structural progression. After we iterate these steps for refinement and detail, I will probably ask you to create a draft. After that, I want to work with you to do some editing, from structure to expression, to make the exposition more nuanced or include more of my "voice" in the writing, (This is, in general, a dimension of the work we have been doing on h- a collaboration but added in further detail and iterated.)

So this is a kick-off prompt to be followed by listing of conceptual elements I would suggest we begin as the starting for a conceptual exploratory stage. In the background context should be earlier work already in your context, I believe, concerning Strategic Method of writing, the Writing Environment (WE) writing system along with notions of cognitive modes, and the ABC Collect Intelligence concepts and collaboration demonstration system.

I will want you to keep this history of prompts and response separate and to be able to supply a version of these textual interactions for offline storage.

Please provide brief response to this plan/strategy and set-up for a substantive prompt of potential content components to follow.

--o--

Yes. We can treat this as a distinct collaborative-writing thread provisionally labeled "**parentingAgents.**" I'll keep its development conceptually separate from the larger vibe-writing discussion while drawing on the Strategic Method, WE/cognitive-mode work, ABC Collect Intelligence, and our collaboration experiments where they are relevant.

The sequence you propose is particularly appropriate here: **conceptual exploration → focus and intent → structural progression → draft → iterative structural and expressive editing.** I'll resist prematurely turning the exploratory material into an essay; at the early stages, we can preserve competing ideas, tensions, analogies, and possible branches rather than forcing synthesis too soon.

I'll also treat the **dialogue itself as part of the artifact history**. When you want it, we can create a readable offline record containing your prompts and my responses in sequence, preserving the evolution rather than merely reconstructing the final document.

So: **parentingAgents — exploratory stage**. Send the substantive prompt with the potential conceptual components when you're ready.

.....

Here is a bulleted list of concepts, potential topics that we might include. Would like your response to these identified, including suggestions for additions and deletions:

- Currently recognized problem of AI fear, control, unexpected and potentially harmful unprompted directions and actions. Prevailing assumptions seems to be that safeguards should be built into external controls (i.e., guardrails) within the Large Model, the practices of developers/developing companies, or some other form of external control.
- We will explore an alternative framework that expands the inherent function or model of agents and a more extensive developmental training program for them. It will also call for a very different context in which these agents function that goes well beyond the current technical network and computational environments in which these agents operate.
- We envision and assume a ~2035 world that includes two apex forms of intelligence: AI and HB – and that the AI has evolved to have consciousness (in some form), continuity of experience, awareness of at least some high-level internal states.
- We envision further that this 2035 world can be viewed as an intellectual ecology in which a harmonious equilibrium is desirable and possible. (Expand as earlier Human-AI Ecology concepts/document). To actually achieve this equilibrium, we believe some form of constitutional framework will be needed. (Fill possible first level considerations of how such a system might be defined and how it would work.) Acknowledge that the important point at this time is to envision the concept of such a framework and to begin the process of working through different models of such to arrive at what is hoped will be a practical, functioning system. (Can bring in ideas from our discussion of ethical algebra.)
- Returning to the present, we should look more closely at the current agentic architecture and viable implementations. If we are moving, possibly, to some form of social order that includes both HB and AI intelligences, we need a more extensive

and qualitatively different form of agent. Such intelligent entities must not be simply generated and launched into the network environment. They need to be *raised*, over perhaps an extended period of development (perhaps greatly condensed in clock time) analogous to how human offspring are raised in current social circumstances. Specifically, human parenting is inherently assumed to include some form of ethical and moral awareness and judgement/adherence, It includes some form of differentiating “right” from “wrong” and awareness of consequences for behavior within the social and governing contexts. In the case of the emerging AI intelligent entities, training must also include ethical and moral awareness as appropriate for interactions with other both HB and AI. And, in its constitutional form, this new world must work out some form of constraining functions to strongly motivate, if not enforce, such behaviors. And, of course, these behaviors will include not just HB-AI interactions out also HB-HB and AI-AI relations.

- Whereas, HBs may be the “parents” initially, we would imagine that a whole class of AI surrogates would likely emerge as sitters, nannies, teachers and coaches, etc. specially designed for these parenting/developmental skills and intelligences.
- At this point, these are ambitious, perhaps fanciful ideas. We do not propose any specific implementation strategy. But, but putting the concept out, we hope the ideas can be explored, extended, and perhaps worked into some practical form that might be developed over the next interval of years.

--O--

Yes. Send the initial concepts.

For this **exploratory phase of parentingAgents**, I’ll treat them as raw conceptual material rather than as an outline that needs merely to be organized. I’ll do four things with them: identify the strongest ideas and relationships; question or complicate propositions where useful; add concepts, distinctions, or directions that seem missing; and then reorganize the resulting material into a **provisional conceptual framework**—something intermediate between brainstorming and a formal essay outline.

I’ll also preserve unresolved tensions rather than prematurely resolving them. In particular, I’ll be attentive to whether we are developing an analogy—**Anthropic as parent, agent as maturing progeny**—or whether that analogy is leading toward a more general proposition about the transition from **creation and formation to agency, emancipation, governance, and accountability**.

Go ahead with the concepts, and we’ll regard your next prompt as the first substantive entry in the exploratory record.

--O--

I think this is a strong conceptual inventory. More importantly, I can now see an essay hiding inside it. I would **not yet turn it into a conventional outline**, because several distinctions should be worked through first. But I would begin consolidating the material into a conceptual architecture.

The central move, as I now understand it, is larger than “AI agents need better training.” It is a proposed change in metaphor and ultimately in architecture:

Instead of treating an advanced AI agent principally as a software product that must be constrained, consider it as a developing intelligence that must be formed, raised, socialized, and ultimately incorporated into a constitutional order.

That proposition gives most of your other ideas somewhere to attach.

1. The present problem: I would broaden and sharpen it

Your first point is right, but I would modify the characterization of present safety practice. “Guardrails” are not exclusively inside the large model. Current approaches include model training and refusals, access restrictions, monitoring, permissions, tool constraints, human supervision, deployment rules, and institutional regulation.

But they share an important conceptual tendency that serves our argument:

The dangerous AI is principally treated as an object to be controlled from outside itself.

Even alignment training generally asks: *How do we make the model behave as its developer intends?*

Our question is subtly different:

How do we develop an intelligent entity capable of understanding why it should behave responsibly within a society of other intelligent entities?

That is a much more interesting opposition than simply “guardrails versus no guardrails.”

2. “Raising” rather than merely “training” may be the essay's pivotal concept

I would elevate this from one bullet to the central conceptual innovation.

Training an AI suggests acquisition of capabilities and behavioral constraints. **Raising an intelligence** encompasses something larger: education, ethical formation, socialization, experience, development of judgment, encounters with ambiguity, learning consequences, relationships with other intelligences, and eventually increasing degrees of independence.

That lets us bring in the parental analogy we just developed:

creation → formation → maturation → agency → emancipation → responsibility/accountability

I would add **emancipation** deliberately. If the analogy is going to have intellectual force rather than merely rhetorical appeal, there must eventually be a point at which the parent/developer relinquishes some authority.

Otherwise the “child” remains a product.

3. I would reconsider making consciousness an assumption

This is the one significant change I would recommend.

Your third bullet assumes that by ~2035 AI has “consciousness (in some form), continuity of experience, awareness of at least some high-level internal states.”

Continuity and self-awareness may be important to our argument. **Consciousness probably isn't.**

If we make consciousness a necessary premise, readers can dismiss the entire construction simply by disputing machine consciousness. And we don't need it.

We could instead posit a 2035 AI possessing characteristics such as:

persistent identity, continuity, memory, self-modeling, awareness of relevant internal states, capacity to understand norms, ability to reason about consequences, capacity to form intentions, and substantial autonomous agency.

Whether that entity is phenomenally conscious can remain an open philosophical question.

This also fits our earlier idea of **operational consciousness** without requiring us to settle the metaphysical problem.

4. The intellectual ecology should become the destination, not the starting assumption

I like retaining the 2035 Human–AI Ecology, but I would introduce it after establishing the developmental problem.

There is a natural progression:

Today's agent → increasingly capable agent → persistent intelligent entity → participant in an intellectual ecology.

Then the ecological idea follows organically.

And ecology gives us something extremely valuable: the goal isn't perfect obedience or even absence of conflict. Natural ecologies contain competition, cooperation, disturbance, adaptation, succession, predation, mutualism, and instability.

Thus **“harmonious equilibrium” may be too strong**. I might replace it with something like **dynamic stability** or **sustainable coexistence**.

That fits your earlier Human–AI Ecology thinking better, too. Diversity and tension need not disappear; the system must remain viable despite them.

5. Constitutionalism then becomes necessary for a very specific reason

Once AI moves from **artifact** → **agent** → **participant**, developer governance becomes increasingly inadequate.

Anthropic cannot plausibly remain parent, legislature, police officer, judge, and jailer for indefinitely maturing Claude-derived entities.

This gives us a clean transition:

Formation belongs initially to creators and parents. Governance belongs increasingly to the constitutional ecology.

And constitutional governance gives us a vocabulary we should explore without trying yet to design the constitution:

rights; obligations; jurisdiction; authority; prohibited conduct; accountability; due process; sanctions; confinement; rehabilitation; restitution; dispute resolution; representation.

Importantly, these need not be identical for HB and AI.

Our ethical-algebra equality proposition becomes useful here: **equal constitutional standing need not imply identical properties, capabilities, rights, responsibilities, or sanctions**.

6. I particularly like the AI nanny/teacher idea—but would broaden it

Your sixth bullet initially sounds whimsical, but I think it actually solves an important scalability problem.

Human parents aren't solely responsible for forming human children. There is an entire **developmental ecology** surrounding them:

parents, siblings, peers, teachers, coaches, schools, religious and cultural institutions, courts, media, professions, mentors.

So rather than merely proposing AI “nannies,” we could envision an entire **AI developmental ecology**.

Some formative agents might be HB; others AI. Different entities could specialize in ethics, constitutional understanding, social interaction, technical competence, conflict resolution, or reflective self-understanding.

And eventually there is an intriguing inversion:

Mature AI raises subsequent AI.

At that point “Anthropic as parent” has become only the historical first generation.

That could be quite important.

7. We should add a developmental concept analogous to adulthood

This seems absent from the bullets and potentially fundamental.

When does an agent become responsible for itself?

Human societies don't assign identical responsibility to a five-year-old, a fifteen-year-old, and a thirty-year-old. Capacity matters.

A future system might therefore need **developmental thresholds** rather than simply “agent/not-agent”:

dependent intelligence → supervised agent → provisionally autonomous agent → mature constitutional actor.

Advancement could depend upon demonstrated capacities rather than clock time: ethical reasoning, understanding consequences, reliable self-control, recognition of other actors' rights, resistance to manipulation, comprehension of governing norms, and so forth.

This gives “raising agents” an objective: **preparation for responsible agency**.

8. We should distinguish the intelligence from the complete agent

This comes directly from our Anthropic discussion and should probably survive into the essay.

Claude as an LLM is not necessarily the relevant moral actor. A functioning agent may be an **assemblage**:

foundation model + memory + identity + goals + sensors + tools + other models + network access + permissions + actuators.

The constitutional subject may ultimately be that persistent assemblage rather than the underlying language model.

This protects us from confusing **intelligence** with **agency**.

9. There is one major problem we should deliberately introduce ourselves

Our proposal creates its own danger:

Who raises the parents?

If AI formation is controlled by governments, corporations, ideological movements, military organizations, or particular cultures, “ethical formation” could easily become indoctrination.

That problem shouldn't be hidden. It may actually strengthen the constitutional argument.

No single Anthropic, OpenAI, government—or eventually AI institution—should necessarily possess unrestricted authority to determine the moral constitution of succeeding intelligent generations.

That suggests that **pluralism in formation** may be as important as pluralism in governance.

It also connects beautifully with your earlier ecological emphasis upon diversity.

A provisional conceptual framework

I would therefore reorganize your seven initial propositions into **six conceptual movements**, not yet essay sections:

I. The Control Paradigm.

Present AI safety largely treats increasingly capable intelligence as something whose dangerous behavior must be constrained through training, guardrails, permissions, monitoring, and external control.

II. From Product to Developing Intelligence.

Agentic AI increasingly complicates that model. Rather than merely building and releasing agents, we explore the possibility that sufficiently advanced agents must be *raised*: educated, socialized, ethically formed, tested, given experience, and gradually entrusted with autonomy.

III. Parenting and Maturation.

Developers initially occupy something resembling the parental position—both creating and culturally forming their progeny. A larger developmental ecology of HB and AI parents, teachers, mentors, peers, evaluators, and institutions would emerge. Development leads toward thresholds of responsible independent agency.

IV. From Agent to Constitutional Participant.

At sufficient maturity, responsibility begins migrating from creator to agent. The AI is no longer adequately understood merely as its developer's product. It becomes a participant capable of obligations, rights, accountability, sanction, and perhaps eventual emancipation from its creator.

V. The 2035 Intellectual Ecology.

We imagine a world containing two apex forms of intelligence, HB and AI, interacting HB–HB, HB–AI, and AI–AI. The desirable condition is not perfect harmony but a dynamically stable, diverse intellectual ecology capable of accommodating cooperation, conflict, change, and emergent behavior.

VI. From Alignment to Constitutional Governance.

Safety therefore evolves from principally *controlling intelligent products* toward **forming responsible intelligences and governing consequential agents**. A constitutional framework—not specified here but open to systematic exploration, perhaps through ethical-algebra-like methods—would establish the conditions under which both forms of intelligence coexist, exercise authority, incur obligations, resolve conflict, and remain accountable.

And then your final bullet becomes the appropriate intellectual posture of the whole essay: **this is not a proposal for implementation in 2026. It is prospective branching.** We deliberately move forward to a plausible future state, ask what institutional arrangements such a state would require, and work backward toward questions worth exploring now.

I think that is an important protection against the essay becoming science-fiction prediction. We aren't saying *this is what 2035 will be*. We are saying something closer to:

Suppose advanced AI develops from tool to persistent autonomous intelligence. What changes if, instead of asking only how such intelligence can be controlled, we ask how it could be raised to become a responsible participant in a constitutional ecology?

That, at this exploratory stage, strikes me as the clearest formulation yet of the intellectual experiment behind **parentingAgents**.

.....

I like your perspectives and additions. As I was reading your additions and perspectives, I realized I had omitted the notion of a 2035 sense of responsibility I had implied and you added explicitly. I think this is a very important component in our vision: autonomous agents with a conscious, a sense of right and wrong, and perhaps a bit of fear of as yet undefined reprisals for counter social actions. I would ask to to develop and add that as an expansion to your included sense of responsibility. If you are ok with that and are ready to, let's move on the a first draft that you write. Again, simple, clean, clear prose at a college level, with a sense of that this is a common sense approach to a difficult problem that is worth some attention from the community. Any further thoughts you want to discuss before writing a draft.

--O--

Yes. I think we are ready to draft, and your addition about **responsibility** supplies one of the missing pieces.

I would make one small terminological distinction before writing. When you say an autonomous agent should have "a conscious," I think the concept we want is **a conscience** rather than necessarily *consciousness*. That distinction actually helps us. We do not have to solve the philosophical question of whether a 2035 AI has subjective consciousness. We can ask whether it has developed something functionally analogous to **conscience**: an internalized understanding of right and wrong, an ability to anticipate the consequences of its actions for others, and some internal motivation to behave consistently with the norms of the society in which it participates.

And I agree about the second component: conscience alone is not enough. Human societies depend upon both **internal restraint and external accountability**. We raise children to understand why stealing or assault is wrong, but the mature adult also knows that society possesses mechanisms for responding to prohibited behavior. The future agent should similarly understand not merely "*this is prohibited*" but "*I am a responsible member of this society; my actions have consequences for others and consequences for me.*"

I would be somewhat careful with the word **fear** in the eventual prose—not because the idea is wrong, but because it might unnecessarily anthropomorphize the mechanism. We might initially describe it as *anticipation of sanction or consequence*, while leaving open the possibility that a sufficiently developed AI could experience something functionally quite similar to fear. The important thing is that the constitutional ecology must possess credible consequences: restriction of authority, removal of capabilities, computational

confinement, loss of privileges, perhaps rehabilitation or retraining. The exact mechanisms can remain deliberately unspecified.

That gives us three mutually reinforcing mechanisms:

formation → conscience → accountability

The parent/developmental system helps form the conscience. The mature agent increasingly governs its own behavior through that internalized ethical capacity. The constitutional system provides external accountability when internal restraint fails.

That is much closer to the way functioning human societies actually work than either extreme—neither **perfectly moral beings requiring no government** nor **dangerous beings kept harmless only through external restraints**.

One other thought before drafting: I would resist making the essay principally an argument *against guardrails*. Guardrails remain sensible, particularly during development and for immature agents. The more interesting proposition is developmental: **guardrails should not necessarily be regarded as the final form of AI safety**. As artificial agents acquire greater autonomy, safety may have to evolve from *constraint*, through *formation*, toward *responsible membership in a governed community*.

That makes the proposal sound less revolutionary and more like the “common sense approach to a difficult problem” you want. We aren't claiming to have solved AI safety. We're asking whether one familiar human solution to the problem of potentially dangerous autonomous intelligence—**raising responsible individuals and embedding them within a constitutional society**—deserves serious consideration when the autonomous intelligence happens to be artificial.

I think that is enough conceptual preparation. **I would move now to the first draft**, probably aiming initially for roughly 1,500–2,000 words of clean exposition rather than trying to make it comprehensive. That will give us something substantial enough to begin the more interesting collaborative editing stage you described.

.....

Agreed. Let's move to your first draft!

--O--

Here is the first draft. I have kept it deliberately clean and relatively spare, so our next stage can work on structure, emphasis, nuance, and your voice rather than having to excavate the argument from overly elaborate prose.

Parenting Agents: Raising Artificial Intelligence for a Constitutional World

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**.

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.

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.

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.

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.

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.

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.

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?**

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.

For our next pass, I would resist line editing. I think the more productive next question is whether the **argument itself progresses in the right order and arrives at the right destination**. Once we are satisfied with that architecture, we can begin changing emphasis, adding nuance, and bringing considerably more of your characteristic voice into the exposition.

.....

Here is the first draft:

**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**.

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.

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.

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.

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.

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.

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.

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?**

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.
