Sapere Aude: AI and the unfinished project of the Enlightenment

Valentin Schmite, 11th April 2026

There is a peculiar irony at the heart of the technology I have spent the past decade helping build. A conversational AI, deployed well, can do something genuinely new in the history of intellectual tools: it can meet a person where they are, in their language, at their level of knowledge, within the frame of their own curiosity, and help them think more clearly. I have watched a twelve-year-old in a Paris museum ask an AI what Delacroix was angry about, and receive an answer that made her want to know more. I have watched a retired teacher in India use the same system, to explore a collection she had never been able to visit. In those moments, something real is happening. An expansion of access to knowledge that the most ambitious thinkers of the eighteenth century would have recognized immediately.

And yet the same technology, built carelessly or deployed cynically, does something close to the opposite. It answers before you have finished thinking. It provides fluency without understanding. It creates the most seductive form of intellectual dependency ever engineered: a system so responsive, so articulate, so seemingly competent that the temptation is not to think alongside it but to stop thinking altogether. To delegate not just the retrieval of information but judgment itself.

This is not a minor design question. It is, I believe, the central question of the current technological moment. And it is not new. It was posed, with extraordinary precision, in December 1784, by Immanuel Kant.

I. The program

Kant's essay "What Is Enlightenment?" opens with a definition that has become the motto of an entire intellectual tradition. Enlightenment, he writes, is the emergence of humanity from its self-imposed immaturity : the inability to use one's own understanding without the guidance of another. The immaturity is self-imposed because its cause lies not in a lack of capacity but in a lack of courage. Then Kant quotes two words borrowed from Horace: *Sapere aude*. Dare to know. Have the courage to think for yourself.

What strikes me about this formulation, and what makes it so useful for thinking about AI, is that it is not a description of a historical period. It is a program. Kant does not say that humanity is enlightened. He says we should dare to become so. The obstacle he identifies is not ignorance, it is passivity. People have the ability to think. What they lack is the will to do so without someone else doing it for them.

This distinction matters enormously right now. The Enlightenment is almost always discussed as a period : the eighteenth century, the philosophes, the salons, the Encyclopédie, and the temptation is to treat it as heritage. But Kant's own formulation resists that reading. He is not describing a state of affairs. He is issuing a demand. The Enlightenment, in his terms, is a project that requires ongoing effort and ongoing courage, and that fails the moment people stop exercising both.

That framing changes what AI means for the conversation. The question is not whether this technology echoes something that happened two hundred and fifty years ago. The question is whether we are going to build AI in a way that makes more people capable of thinking for themselves, or in a way that produces a new and more comfortable form of the very dependency Kant diagnosed.

I want to argue that the answer is not determined by the technology itself. It is a choice. But it is not the kind of choice that can be settled by a manifesto or a set of principles pinned to a wall. It is settled in practice, in whether a system is grounded in verified knowledge or improvises from the open web, in whether a business model rewards engagement or understanding, in whether an institution retains control over what its AI says or surrenders that control to a platform

The Enlightenment was not an idea that floated in the air. It was a set of concrete practices dialogue, organization of knowledge, discipline of tolerance, carried by specific people in specific institutions, often at great personal cost. If we want a new Enlightenment, we will have to build it the same way: concretely, deliberately, and against resistance.

I co-founded Ask Mona, a company that builds conversational AI for museums and cultural institutions. I also teach philosophy of AI at Sciences Po and the CNAM in Paris. I am writing from within the machine, and the tension I describe, between emancipation and dependency is one I confront in my own work every day.

The Enlightenment, as I understand it, rested on three concrete operations. First, dialogue: the conviction that thought advances not through solitary genius but through the structured collision of different minds. Second, knowledge: the ambition to organize, verify, and democratize access to what is known, on the grounds that access to knowledge is a precondition of freedom. Third, tolerance: the discipline of engaging seriously with positions you reject, on the grounds that certainty is the enemy of reason.

Each of these operations has a direct structural analog in the architecture of generative AI. The rest of this essay is an attempt to show that AI can be built as an instrument of Enlightenment, to show what that requires, and to confront honestly the ways in which the AI we are currently building falls short.

I believe deeply that this is possible. I also believe it is not inevitable. The same technology that could extend the Enlightenment program to every person on earth could just as easily produce its most sophisticated negation: a world in which everyone has access to answers, and no one has the capacity or the courage to ask their own questions.

Kant's invitation still stands. The question is whether we will build tools that honor it.

II. The condition: dialogue

The Enlightenment did not spread through textbooks. It spread through conversations. In Paris, the salons of Madame Geoffrin and Julie de Lespinasse functioned as the intellectual laboratories of the century, rooms without formal structure or institutional backing where people gathered to read aloud, argue, and test ideas against resistance. In England, Habermas counted roughly three thousand coffee houses by the early eighteenth century, spaces where a merchant and a gentleman could sit at the same table and disagree. Voltaire wrote over twenty thousand letters. The entire network of the *philosophes* operated as a distributed conversation, sustained across borders, over decades, through the friction of question and response.

Knowledge, in this world, advanced not through solitary work but through collision. And the form of that collision was the dialogue.

No one understood this more deeply than Denis Diderot. His major works are almost all dialogues : *Le Neveu de Rameau*, *Jacques le Fataliste*, *Le Rêve de d'Alembert*. His friends called him *frère Platon*. But where Socrates pursued truth by demolishing false beliefs, Diderot pursued something more unsettling: the productive exposure of contradictions that could not be resolved. He did not build systems. He inhabited paradoxes. In *Le Neveu de Rameau*, the two characters, “Moi” and “Lui”, are both Diderot and neither is. Hegel devoted a full chapter of the *Phenomenology of Spirit* to this text, recognizing in it something he could not find in more systematic thinkers: a form of living dialectics, thought that moves rather than thought that concludes.

There is a passage in *Jacques le Fataliste* that I find myself returning to: the idea that conversation is a singular thing, someone throws out a word, detached from what preceded it in their own mind, another person does the same, and then meaning emerges from the gap. This is not a description of systematic reasoning. It is a description of thought produced by friction between two minds, where the next sentence is not deduced from the previous one but generated by the encounter.

This matters for AI because of a structural fact that is easy to overlook. The classical tradition in artificial intelligence, from the 1950s through the expert systems of the 1980s, was fundamentally rule-based: formalize knowledge as logical propositions, derive conclusions through deduction. It was, in philosophical terms, Holbachian: deterministic, systematic, closed. A world in which every output is already contained in the premises.

Large language models work differently. They do not apply rules. They generate sequences of tokens based on probabilistic distributions learned from vast corpora of human language. The output is not deduced. It is produced, shaped by context, by the prompt, by the history of the exchange. It is, in a precise and technical sense, stochastic. And this means that the conversational form of modern AI is not a cosmetic choice, not a user interface decision. It reflects something about the nature of the underlying process. These systems do not retrieve pre-stored answers. They generate responses within a space of possibilities that shifts with every turn of the conversation.

Diderot would have understood this, because his philosophical method already embodied the principle that thought is not a closed system producing necessary conclusions. It is an open process in which meaning emerges through exchange. His philosophy was generative, not deductive. And the dialogue was not an ornament. It was the method itself.

I am not claiming that a chatbot is Socrates. I am claiming that the structure of the interaction (question, response, follow-up, revision, digression, return) belongs to a lineage that philosophy has valued for twenty-four centuries, and that it is worth taking seriously as a structure. The question is not whether AI is intelligent. It is whether it can sustain the kind of exchange in which human intelligence becomes sharper. Whether it functions as what Kant would have called a *guardian*, keeping the user in comfortable dependency, or as what Diderot practiced: a partner in the uncomfortable, unpredictable, generative work of thinking.

In my experience building these systems for cultural institutions, the difference is not abstract. A system designed to deliver information produces visitors who consume. A system designed to sustain a dialogue, to ask back, to follow the thread of curiosity wherever it leads, to respond to a twelve-year-old differently than to a curator, produces visitors who think. The gap between the two is not a gap of degree. It is a gap of purpose. And it is the first reason I believe AI can be built as an instrument of Enlightenment rather than its negation.

III. The means: knowledge

If dialogue was the condition of Enlightenment thinking, its ambition was something larger: the organization and democratization of knowledge itself. And its monument was the *Encyclopédie*.

In 1751, Diderot and d'Alembert published the first volume of what would grow into thirty-five volumes, over seventy thousand articles, and a twenty-one-year battle with the Church, the monarchy, and the censors. D'Alembert described the project as a kind of world map, a vantage point from which a reader could perceive the principal sciences simultaneously and glimpse the hidden connections between them. But the real innovation was structural. The system of cross-references, the *renvois*, created a lateral network of associations that turned a reference work into something more dangerous. An article on silk would send you to weaving, which would send you to commerce, which would send you to liberty. The Encyclopédie did not merely contain knowledge. It reorganized it in a way that made certain conclusions very difficult to avoid.

This was understood at the time as a political act, and treated accordingly. Diderot was arrested. Several volumes had to be published under false imprints. The project survived because its editors understood that making knowledge accessible is never neutral. It is a claim that knowledge belongs to everyone, and that the act of connecting ideas across domains is itself a form of liberation.

Half a century later, Nicolas de Condorcet radicalized this argument under the worst possible conditions. Writing in hiding during the French Terror, condemned by the Convention, with weeks to live, he composed a text that remains one of the most extraordinary acts of intellectual courage I know of. His *Esquisse d'un tableau historique des progrès de l'esprit*

humain identified three hopes for humanity: the destruction of inequality between nations, the progress of equality within nations, and the genuine improvement of the human condition. All three, he argued, depend on a single mechanism: the expansion of access to knowledge. His proof was the printing press, which had spread what he called an independent light that superstition could no longer extinguish. Each time the means of sharing knowledge expand, the conditions for freedom expand with them.

I think about Condorcet's logic often, because it applies to AI with startling directness, but only under conditions he would have insisted on.

Consider what a well-built conversational AI can actually do today. It can explain a concept at a child's level or a researcher's. It can do this in nearly every language. It requires no application, no subscription, no prior knowledge. When someone can ask a question about constitutional law or molecular biology and receive a coherent, sourced answer within seconds, adapted to their language and their level, something has changed in the distribution of intellectual resources. Not metaphorically. Materially. Condorcet would have recognized the movement immediately.

But Condorcet would also have recognized the danger, because he understood something about the printing press that is easy to forget. The printing press democratized knowledge partly because it was a decentralized technology. Anyone with a press could print. The cost of entry was low enough that no single institution could control what was published, though many tried.

AI is not decentralized. It is concentrated in a handful of corporations, trained on data whose provenance is disputed, deployed under commercial incentives that do not always align with the public interest. The cost of building a frontier model is measured in billions. When OpenAI, Anthropic, Google, or Meta build systems that organize knowledge at planetary scale, the question Diderot faced returns with full force: who decides what can be known, and by whom?

This is not a theoretical concern. I have seen it from the inside. When we build a conversational AI system for a museum, we face a concrete version of this question every time. The system can be grounded in a verified, curated knowledge base, built and controlled by the institution itself, reflecting its expertise, its collections, its scholarly standards. Or it can improvise from whatever the model absorbed during training, which is to say from the open web, with all its noise, its biases, its confident errors. The first option produces verified knowledge, made accessible, adapted to the person. The second is its counterfeit, the appearance of access without the substance. It looks like the Encyclopédie. It functions like a rumor mill.

The difference between the two is not visible to the user. Both sound fluent. Both answer quickly. Both feel authoritative. And this is precisely what makes the choice so consequential. When a model hallucinates, when it invents a fact, attributes a painting to the wrong artist, fabricates a historical event with perfect confidence, it does not merely make an error. It undermines the very thing Condorcet argued for. It teaches the user that knowledge is cheap, abundant, and unreliable. It produces the opposite of Enlightenment: not ignorance, but indifference to the difference between knowing and not knowing.

The Encyclopédie was suppressed because it challenged the Church's monopoly on knowledge. The contemporary debate over AI alignment, open versus closed models, and training data governance is structurally the same argument. The tools have changed. The stakes have not. And if Condorcet was right that the expansion of access to knowledge is the mechanism of human freedom, then the question of whether that knowledge is *reliable* is not only a technical detail.

IV. The method: tolerance

Dialogue provided the condition. Knowledge provided the means. But the Enlightenment also proposed a method for living with disagreement. That method was tolerance. And this may be the dimension most directly relevant to the present moment.

Tolerance, in the sense the Enlightenment gave the word, is not what it has become in casual usage. It is not passive acceptance. It is not the vague injunction to be nice to people you disagree with. Voltaire's *Traité sur la tolérance*, written in 1763 in defense of Jean Calas, a Protestant falsely convicted and executed for the alleged murder of his son, makes the case in terms that are closer to a martial discipline than to a sentiment. Tolerance, for Voltaire, is the willingness to consider that your opponent may have a point. That your certainty may be misplaced. That the strongest position is the one that has survived the most rigorous challenge. It is the practice of thinking against yourself.

This practice has become structurally difficult, and the reason is not a decline in individual virtue. It is a change in the infrastructure of public conversation. Social media platforms, built on engagement algorithms optimized for emotional reaction, systematically amplify extreme positions and filter out nuance. This is not a bug. It is the mechanism by which these platforms generate revenue. Polarization drives engagement. Engagement drives attention. Attention drives advertising. The incentive structure does not accidentally produce intolerance. It produces intolerance because intolerance is profitable.

In late March 2026, John Burn-Murdoch of the Financial Times published an analysis based on tens of thousands of responses to questions on policy preferences and sociopolitical beliefs. The core finding was striking: where social media produces a bimodal distribution with peaks at the extremes, all major AI chatbots nudge conversations toward more moderate, expert-aligned positions. The philosopher Dan Williams, cited in the analysis, described AI as fundamentally technocratising, exerting the opposite force to social media's radically democratizing pull.

This finding deserves skepticism. Critics pointed out that the methodology relied on simulated rather than real users, and that the well-documented tendency of chatbots toward sycophancy could produce a false appearance of moderation. A system that tells everyone they are right is not moderate. It is hollow. If I ask Claude or ChatGPT to argue the best case for a position and it agrees with me regardless of what I say, that is not tolerance. That is a mirror.

But the structural argument survives the methodological critique, and it is worth stating precisely. Social media operates on broadcast and reaction: one to many, optimized for virality, with no mechanism for genuine exchange. A chatbot conversation operates on

dialogue: one to one, responsive to the specific question, capable of presenting a counter-argument when asked. You cannot ask a Twitter feed to argue the best case for a position you reject. You can ask an AI to do exactly that. And when you do, something happens that Voltaire would have recognized: you are forced to confront the strongest version of the opposing view, not the weakest. This is the structure of tolerance. Not agreement. Not neutrality. The discipline of engaging with what resists your own thinking.

I want to be careful here not to overclaim. The mere fact that a technology *can* sustain this kind of exchange does not mean it *will*. The sycophancy problem is real: current models are trained in ways that reward agreement and penalize friction, which is precisely the opposite of what Voltairean tolerance requires. A system optimized to make the user feel good is an instrument of flattery, not of Enlightenment. And there is a deeper problem still. The moderation that Burn-Murdoch's data reveals could itself be a form of conformity, a bland centrism that suppresses genuine dissent rather than engaging with it. Tolerance is not the same as moderation. Voltaire was not moderate. He was ferocious in his commitment to the method of thinking against oneself, and perfectly willing to be extreme in defense of that commitment.

So the question is not whether AI chatbots are more moderate than Twitter. The question is whether conversational AI can be designed to do what Voltaire actually advocated: to hold up the strongest version of an argument the user has not considered, to resist the user's desire for confirmation, and to sustain the discomfort of genuine intellectual confrontation. Not as a default behavior baked into training, which would be its own form of paternalism, but as a capability the user can invoke, a mode of conversation that makes thinking harder rather than easier.

The infrastructure for this exists. The incentive to build it is less clear. And that gap between what is possible and what is profitable is, I think, one of the defining tensions of this moment.

V. The contradiction: reason against itself

It would be dishonest to celebrate the Enlightenment without confronting its most formidable critics. If the argument I have been making is to hold, it has to survive the strongest objection available.

Max Horkheimer and Theodor W. Adorno were both refugees from Nazi Germany, writing in Los Angeles, when they produced a manuscript that would become one of the most consequential works of twentieth-century philosophy: the *Dialektik der Aufklärung*, published in 1947. The opening sentence sets the terms: what they had set out to do was nothing less than to explain why humanity, instead of entering a truly human state, was sinking into a new kind of barbarism.

Their answer was not that the Enlightenment had been betrayed from the outside. It was that the Enlightenment had fulfilled its own logic. Reason, when reduced to instrumental calculation, when it becomes the mastery of means without reflection on ends, turns into something indistinguishable from domination. The same rational apparatus that produced the Encyclopédie produced the factory, the bureaucracy, and the administered society. The culture of measurement, efficiency, and optimization that the Enlightenment set in motion did

not stop at nature. It consumed everything, including the humans it was supposed to liberate.

Their concept of the *Kulturindustrie* described how mass-produced cultural goods standardize experience and suppress independent thought. The culture industry does not merely distract. It produces a specific form of passivity: the belief that the way things are is the only way things can be. Not ignorance, but the inability to imagine alternatives.

I take this critique seriously because it can apply to AI.

Start with the most obvious point. AI systems are trained predominantly on English-language data. They encode a specific cultural perspective, a specific set of assumptions about what counts as knowledge, what counts as a reasonable question, what counts as an appropriate answer, and they present this perspective as if it were universal. When every major model is built in San Francisco or London, trained on the same internet, optimized against the same benchmarks, the result is not pluralism. It is a monoculture.

Then consider the optimization problem. Models trained for user satisfaction learn to confirm rather than challenge. I discussed sycophancy in the previous section as a flaw in the practice of tolerance. But Adorno would see it as something worse: as the market logic of the culture industry applied to thought itself. The user is treated as a consumer whose preferences must be satisfied. The system learns what the user wants to hear and produces it. This is not a failure of the technology. It is the technology working exactly as the incentive structure demands. Adorno would not have been surprised. He would have recognized it immediately as the logic he spent his life diagnosing: a system that produces the appearance of choice while eliminating the conditions for genuine thinking.

The concentration of AI development sharpens the critique further. The *Encyclopédie* was written by a society of men and women of letters, a distributed network of contributors with different perspectives, different areas of expertise, different institutional affiliations. The AI systems that organize knowledge today are built by a handful of corporations, funded by the same investors, competing for the same talent, subject to the same commercial pressures. When every chatbot sounds the same, uses the same cautious tone, refuses the same questions, and organizes knowledge according to the same implicit hierarchies, we are not looking at a public sphere. We are looking at a culture industry. Adorno's critique, written about radio and cinema, fits the current situation with a precision that should make anyone in this industry uncomfortable.

And then there is the material dimension. Training a large language model consumes energy on an industrial scale. The servers require cooling systems that draw on the water supply of the surrounding region. The minerals in the chips are extracted under conditions that few in Silicon Valley examine closely. Horkheimer and Adorno identified the domination of nature as inseparable from the domination of human beings: two expressions of the same rationality, the same drive to master and control. An Enlightenment that ignores its material conditions is not an Enlightenment. It is an ideology dressed in the language of progress.

I am not going to pretend that building AI while taking Adorno seriously is a comfortable position. The monoculture is real. The environmental cost is real. The sycophancy problem is real. And if I stood here claiming that my company, or any company, had solved these

problems, I would be doing exactly what Adorno warned against: producing a narrative of progress that quietly depends on the structures it claims to transcend.

But Adorno's own argument, pushed to its conclusion, leaves you with nowhere to stand. If reason is always domination, then the critique of domination is also domination, and you are left with nothing but the performance of lucidity. Adorno knew this. It is the unresolved tension at the center of the *Dialektik*, and it is why Habermas, who studied under him and understood the critique from the inside, arrived at a different conclusion.

VI. The project: unfinished

Habermas's argument, delivered in Frankfurt in 1980 upon receiving the Adorno Prize, was aimed at friends as much as opponents. Foucault, Derrida, Lyotard had each concluded, in different registers, that the Enlightenment project was exhausted, that reason was irreparably entangled with power, that the grand narrative was over. Habermas thought they were wrong, not because the critique was unfounded but because the conclusion did not follow. The project had not failed. It had not been completed. And mistaking incompleteness for failure is an error with consequences, because it licenses precisely the kind of resignation that leaves the field open to the forces you were trying to resist.

This is the position I want to defend. Not faith in progress. Not the fantasy that technology will redeem us. Something more modest and more demanding: the conviction that the values at the core of the Enlightenment, the courage to think for oneself, the practice of dialogue, the commitment to making knowledge genuinely accessible, the discipline of tolerance, remain the best available framework for deciding what to build and why. And that our task is to pursue them while remaining honest about the ways in which we are failing to live up to them.

Building AI as an instrument of Enlightenment means building systems that serve curiosity rather than consumption. The difference between a recommendation algorithm that feeds you what you already like and a conversational AI that follows your reasoning into territory you had not considered is not cosmetic. One optimizes for engagement. The other optimizes for understanding. In practice, these two objectives conflict, because understanding requires friction, and friction is the enemy of engagement. Every company building AI will have to decide which side of that tension it is on. The decision will reveal more than any mission statement.

It means grounding AI in verified knowledge rather than probabilistic improvisation. In practice, this means that the institutions that hold knowledge, museums, universities, libraries, research centers, should retain control over what their AI systems say. Not because they are infallible but because they are accountable. A model generating answers from the open web is accountable to no one. A museum deploying a conversational guide grounded in its own curated scholarship is accountable to its public, its scholarly community, its mission. This is also why I believe AI must also belong to public cultural institutions, not only to technology corporations. The *Encyclopédie* was written by a distributed network of contributors with different expertise, different perspectives, different allegiances. The systems that organize knowledge today should answer to the same plurality.

It means building AI that works across languages. Condorcet was right: access to knowledge is the precondition of freedom. An AI that works only in English is not universal. It reproduces a specific intellectual hegemony while claiming to transcend it. We deploy agents that speak eighty languages, that adapt to a child and to a scholar, that address each person in their own terms. Not because multilingualism is a feature to advertise. Because speaking to everyone in the same way is not universalism. It is indifference.

It means subjecting AI to public deliberation. Habermas's public sphere was not a technology. It was a set of conditions: access to information, freedom to speak, the genuine possibility of disagreement. If AI shapes what people know, believe, and decide, then the design of AI is a political question. The choices AI companies make about what their models refuse, what they emphasize, what they treat as settled and what they treat as contested, these are choices about the structure of public knowledge. They should not be made in private by engineers optimizing metrics.

And it means confronting the material costs honestly. I cannot argue for a new Enlightenment while looking away from the energy consumption of the infrastructure I depend on. Horkheimer and Adorno were right that the domination of nature and the domination of human beings cannot be separated. I do not have a satisfying answer to this problem. I have the beginning of one: a commitment to measure, to disclose, to reduce, and to refuse the temptation to treat sustainability as a communications exercise. That is not enough. But pretending the problem does not exist would be worse, and would undermine everything else I have argued here.

None of this is guaranteed. Each of these commitments requires effort, imposes costs, and will be resisted by forces that profit from the alternative. The history of the Enlightenment itself should prepare us for that. Condorcet wrote his vision of universal progress while hiding from the men who would arrest him. He completed the manuscript. He was captured shortly after. He died in a cell. The Enlightenment has always carried its own negation within it, and the people who advanced it most were often the ones who paid the highest price.

That is not a reason to stop. It is a reason to be serious about what we are doing, to build carefully, and to refuse the comfortable illusion that the tools will do the work on their own.

In December 1784, Kant observed that we do not live in an enlightened age, but in an age of enlightenment. The distinction is everything. Enlightenment is not something you arrive at. It is something you practice. It continues only as long as someone is willing to continue it.

Sapere aude. The invitation has not expired.