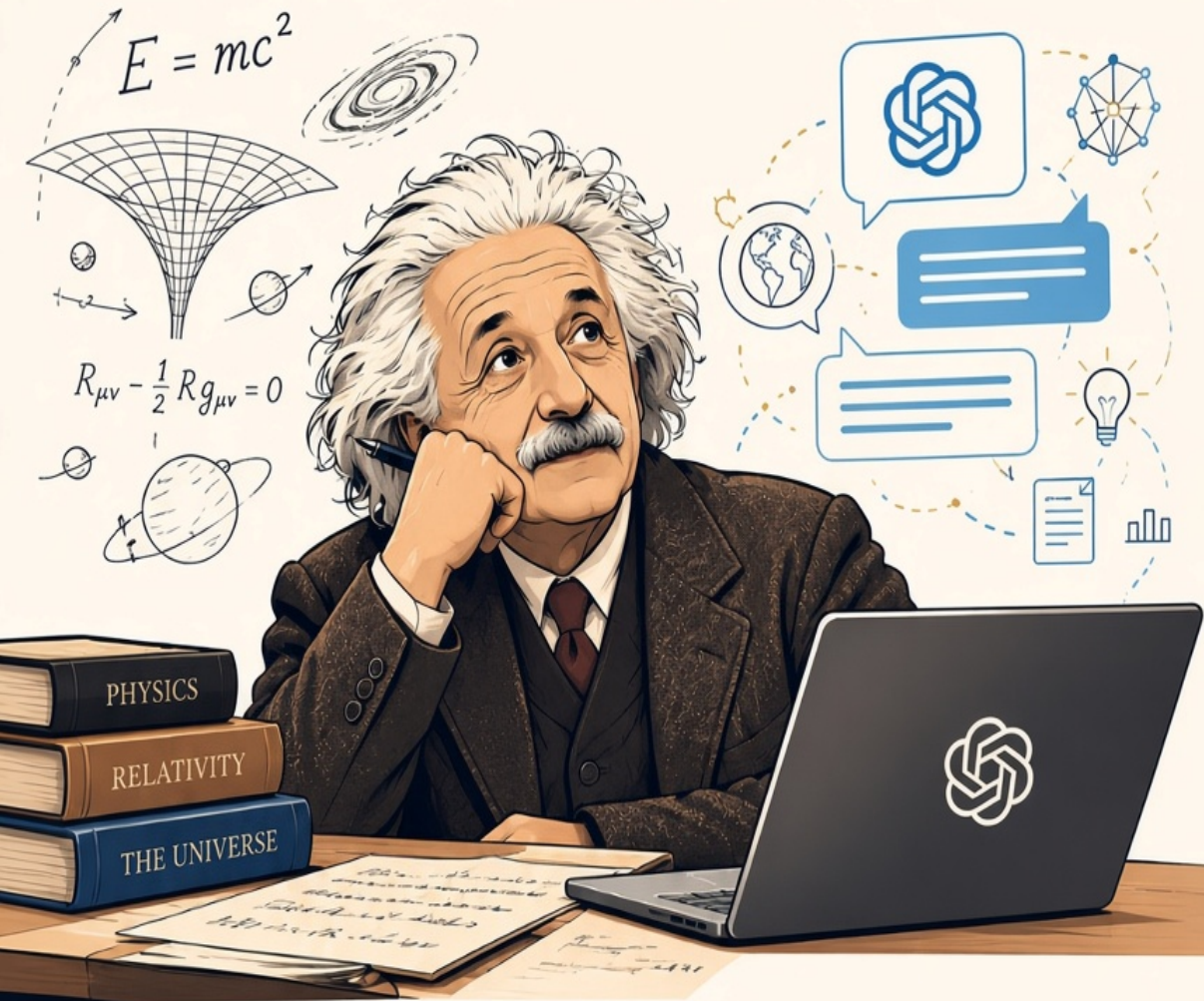


IF EINSTEIN HAD HAD CHATGPT

What artificial intelligence could learn from the way we think —
and what we are teaching it about the world



FRANCK MEVENGUE

INTERIOR TITLE PAGE

IF EINSTEIN HAD HAD CHATGPT

What artificial intelligence could learn
from the way we think — and what we are
teaching it about the world

FRANCK MEVENGUE

Transparency note

A book about AI, written with AI

Before anything else, a small experiment is required: this book is about artificial intelligence... and it was itself co-written with ChatGPT. It would have been a pity to discuss the experiment without trying it.

At its origin, this project is also an editorial experiment: exploring how far artificial intelligence can help us design, research, structure and write a relevant book or article, without giving up intellectual rigor or human creativity.

The ideas defended, the viewpoints and the orientations of this work remain my responsibility as author. ChatGPT accompanied me in documentary research, the exploration of ideas, the structuring of the manuscript and the drafting work. Important sources and claims are meant to be examined and verified by me. A convincing AI answer never constitutes, by itself, a proof.

I want this collaboration to be transparent. We increasingly accept that AI takes part in software design, data analysis or business productivity. Why should art, writing and creation be condemned to remain outside that exploration?

The tool does not necessarily replace the author: it can also shift the boundaries of what the author is able to imagine, explore and achieve. One must still know how to ask the right questions, exercise judgment and own the final choices.

After all, if Einstein had had ChatGPT, he might have tested his hypotheses with an artificial intelligence. For my part, I began by testing whether it could help me write a book on that subject.

The loop is closed. Or almost: the sources still remain to be checked.

Full manuscript

Prologue, ten parts and epilogue

THOUGHT EXPERIMENT

2026: Einstein opens ChatGPT

CAN ONE MODEL A THOUGHT?

- 1 · What is a thought?
- 2 · The mind leaves traces
- 3 · Can one know someone without being them?

IF EINSTEIN HAD HAD CHATGPT

- 4 · Einstein in 2026
- 5 · Einstein's personal chatbot
- 6 · The Einstein model
- 7 · Would Einstein have recognized his own model?

HOW AN AI LEARNS

- 8 · From data to representations
- 9 · Foundation models
- 10 · ChatGPT is not a database
- 11 · Context, memory, and parameters

THE HUMAN AS DATA

- 12 · Is every conversation data?
- 13 · Does the researcher feed the AI?
- 14 · Does AI also learn our methods?

AI AND THE WORLD MODEL

- 15 · What is a "world model"?
- 16 · Does AI possess a representation of the world?
- 17 · From millions of traces to a collective representation

AN AI THAT GROWS WITH HUMANITY

- 18 · 2026 → 2030 → 2040 → 2050
- 19 · Does each generation inherit from the previous one?
- 20 · The AI – humanity loop

HUMAN MODELS

- 21 · Can AI learn "who we are"?

-
- 22 · The intellectual twin
 - 23 · A library of scientific personalities

THE PROBLEM OF TRUTH

- 24 · Data $\neq$ thought
- 25 · Correlation $\neq$ understanding
- 26 · Hallucination
- 27 · The mirror bias

THE GREAT PARADOX

- 28 · We teach AI to understand us
- 29 · But AI also learns our models of the world
- 30 · Who learns from whom?

2050: WHAT HAVE WE CREATED?

- 31 · Humanity's artificial memory
- 32 · Artificial collective intelligence
- 33 · If Einstein had really had ChatGPT

EPILOGUE

The mirror

2026: Einstein opens ChatGPT

An impossible scene to pose a real question: what can a machine learn from the traces of a mind?

■ THOUGHT EXPERIMENT

The entire narrative of this prologue is fictional. Albert Einstein died in 1955. The gestures, the screen, the questions, and the reactions described here have no historical value. They serve only to construct a thought experiment.

A thought experiment organizes an imagined scenario in order to explore the consequences of a question or of principles. It does not turn that scenario into a historical event [\[SRC-038\]](#).

The blank window

Imagine a workspace in 2026. On the table, sheets carry equations, arrows, words rewritten twice. A notebook lies open beside a computer. Albert Einstein - not the historical Einstein, but the one in our thought experiment - looks at a conversation window that is still empty.

He might begin with a physics question. He might also submit an unfinished page, ask for an objection, compare two formulations. Nothing authorizes us to decide what he would actually type. The important detail lies elsewhere: each time he used the assistant, he would leave a new trace. A dated question. A copied equation. A reply accepted, ignored, or corrected. Perhaps a series of attempts that would show less a solution than a way of searching.

Suppose now that this assistant does not receive a single question, but an organized set of documents: articles, drafts, notebooks, letters, lectures, and corrections. The edited Einstein archives attest to a considerable written legacy, composed of scientific texts, correspondence, and other documents situated in time [\[SRC-001\]](#) [\[SRC-032\]](#) [\[SRC-036\]](#). Our scenario adds to these historical archives contemporary interactions that are entirely imagined.

The assistant might retrieve a passage, bring two formulations together,

or flag that an idea appears in several forms. It might learn regularities of vocabulary, structure, and argument. If given enough examples, it might perhaps produce a text that resembles, under certain criteria, the texts observed.

But what exactly would it have learned?

The temptation of the shortcut

We are quick to answer: "It would have learned Einstein." The formula is seductive because it compresses an entire chain of operations into three words. It is also dangerous.

The system would not have access to a thought before it was expressed. It would receive inscriptions: characters, digitized images, metadata, perhaps transcriptions. It would not see the moment an intuition appears, the sensation of a difficulty, attention shifting, or the reason a sentence stays in a drawer. Even an immense corpus contains only what has been formulated, preserved, recovered, selected, and made readable.

Saying this does not make traces insignificant. A crossing-out can show that a formulation was rejected. A sequence of drafts can document a transformation. A letter can place a problem back into a conversation. These elements support inferences. They do not erase the distance between the document and the mental activity from which it arose.

That distance is the book's first thread:

an AI does not read a mind directly; it processes accessible traces and can build representations useful for predicting, classifying, inferring, generating, or simulating certain aspects of what has been observed.

The verb *can* is essential. The scope of a representation depends on the data, the architecture, the learning objective, the tools, the context, and the way it is evaluated. A convincing imitation in one situation guarantees neither the model's general fidelity nor an identity with the person.

Four mechanisms behind a single screen

In our fictional scene, Einstein writes in a single window. That visual unity can conceal several mechanisms.

The model first has **parameters**, adjusted during prior training. The ongoing conversation then supplies a **context**, used to compute the present reply. A product may also keep a **persistent memory** and reintroduce it in later exchanges. Finally, a system may perform an **external search** over

documents and place the retrieved passages into the context.

These four elements are not interchangeable. A sentence typed in a conversation does not immediately rewrite, by its mere presence, the parameters of the general model. In the few-shot evaluation described for GPT-3, tasks are performed without a gradient update [SRC-009]. At OpenAI, the supervised fine-tuning documentation describes a different operation: a set of examples is uploaded, then a training job updates the model weights [SRC-039]. A service's policies may provide that some portion of certain data will later be used in an improvement process; that belongs to another temporality, to controls, and to processing choices. ChatGPT's current documentation distinguishes, among other things, memory controls from controls concerning the use of conversations to improve models [SRC-006] [SRC-007]. This description is dated: it characterizes a product and a policy that may evolve, not a universal law of artificial intelligence.

Our imagined Einstein could therefore have before him an assistant whose replies combine several sources of information. To know what the system actually used, the appearance of the reply would not suffice. One would need to document the configuration.

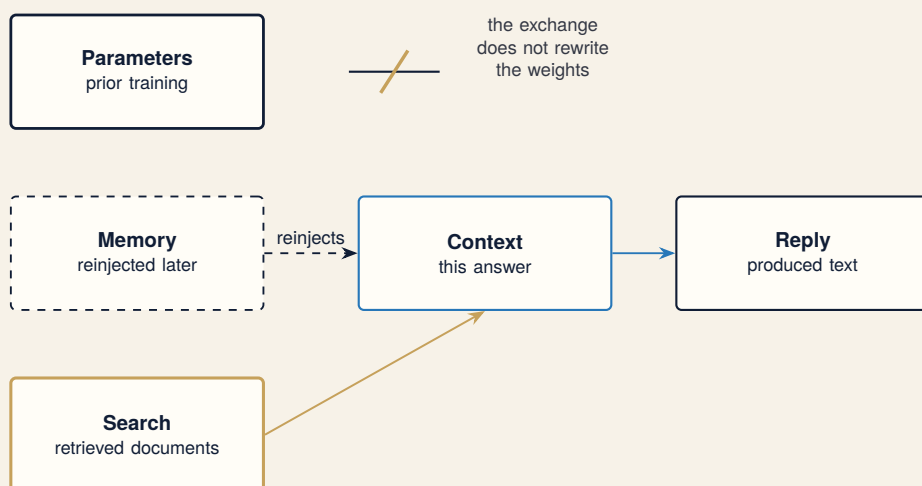


Figure. Parameters, context, product memory, and external search are not the same operation. The exchange feeds the context. It does not update the weights. Explanatory diagram, not an experimental result. [SRC-006] [SRC-007] [SRC-009] [SRC-039], at the registry's review levels, product documentation dated 9 October 2026.

■ ESTABLISHED

Learned parameters, current context, a product's persistent memory, and external document search are distinct mechanisms. A conversation must not be described as an instantaneous modification of the general model.

The double test

Imagine that after months of interaction, the assistant receives a new problem. It proposes a reply. Specialists judge it consistent with certain habits observed in Einstein's writings. What would that success prove?

It would first show a performance: under the test conditions, the system produced something that satisfies defined criteria. If the documents used for evaluation had been held out from the model's construction, the result would be more informative than simply repeating a known passage. If several evaluators, several problems, and several formulations converged, the inference would gain strength.

It would nevertheless remain limited. A text can resemble another for the wrong reasons. A model can exploit superficial cues. Evaluators can share the same expectations. Above all, success would not demonstrate that the system experiences the problem as Einstein would have experienced it, nor that it possesses his identity, his memories, or his consciousness.

The test should therefore be double. One would need to measure what the model reproduces and actively seek where it fails. Which changes of formulation break its coherence? Which claims contradict the documents? Which aspects of the scientific approach are missing? When does it produce a plausible reply with no provenance?

A serious representation is judged not only by its resemblances. It is also judged by the map of its limits.

Two mirrors

The thought experiment begins with an individual, but it cannot remain there. Contemporary systems are trained on sets that gather productions from many people, institutions, languages, disciplines, and periods. Texts, images, code, and measurements become data after operations of collection, filtering, and transformation.

The book's first mirror follows the chain:

individual → expressed thought → traces → data → model of certain aspects of the person.

The second widens it:

humanity → observations and productions → data → learning → representations used to predict and generate.

In both cases, the same warning applies: the map is not the territory. Data are not the totality of what they represent. They can be abundant and yet incomplete, precise and yet selected, useful and yet biased.

Comparing the two mirrors does not assume that a person and the world are modeled in the same way. It serves to pose a shared question: what does a representation keep when it is built from traces, and what does it lose?

What the scene forces us to ask

Return to the blank window. The fictional Einstein of 2026 might close the computer, take up his notebook again, and leave the reply uncommented. That silence, too, would be ambiguous. Would it mean agreement, doubt, fatigue, or the decision to search another way? Without a further trace, we would not know.

That ignorance is not a temporary defect that more computation necessarily erases. It recalls that all knowledge of a person passes through observations and interpretations. Humans already live with this limit when they read a letter, reconstruct a discovery, or try to understand someone close to them. Artificial intelligence does not create this problem. It enlarges it, automates it, and forces us to clarify our criteria.

The chapters that follow will therefore begin before ChatGPT. They will ask what we call a thought, how it leaves traces, and what it means to know someone from their productions. That detour is indispensable. Without it, we would risk confusing the fluency of a reply with access to the mind that inspired it. This caution about the vocabulary used for language models is itself defended in the contemporary literature [\[SRC-019\]](#) [\[SRC-025\]](#).

KEY POINTS

01 The prologue is a thought experiment, not a historical narrative.

02 An assistant processes accessible traces, not a mind directly.

- 03 Parameters, context, memory, and external search must remain distinct.
 - 04 An imitation is validated by tests and by the study of its failures.
 - 05 The individual question opens onto that of world representations built from human traces.
-

Sources used

- **SRC-001 - Einstein Papers Project.** Institutional scope of the edited writings, correspondence, and documents. Consultation of the official presentation and documentary portal.
- **SRC-006 - OpenAI, Data Controls FAQ.** Product documentation consulted on October 9, 2026; dated policy subject to change.
- **SRC-007 - OpenAI, Memory FAQ.** Product documentation consulted on October 9, 2026; description of memory as a personalization mechanism.
- **SRC-009 - Brown et al., few-shot learning.** Notice and abstract examined; used for task performance without a gradient update in the evaluation described.
- **SRC-039 - OpenAI, supervised fine-tuning.** Official page consulted on October 9, 2026; it describes weight updates as the result of a training job, not of an isolated message. The availability of this platform evolves.
- **SRC-019 - Mitchell and Krakauer, debate on language-model understanding.** Notice and abstract verified; perspective article, not a single experimental verdict.
- **SRC-025 - Shanahan, language used about large language models.** Metadata, abstract, and conclusion excerpts examined.
- **SRC-032 - Einstein Papers Project, history of the project.** Institutional page consulted in full; it describes the organization and editing of the papers after 1955.
- **SRC-036 - Collected Papers, volume 5.** Official editorial notice consulted; it describes the scope of more than 500 letters without authorizing any individual citation.

- **SRC-038 - Brown and Fehige, thought experiments.** Academic entry consulted in full; it frames the methodological status of the imagined scenario.

1 · What is a thought?

Before asking a machine to model a thought, one must clarify what that word actually authorizes us to observe.

One word for several phenomena

"I have a thought." The sentence seems to name something simple. It can nonetheless point to a mental image, an inner sentence, a memory, an intention, a comparison, a calculation, an interpreted emotion, or a question still without words. The same term gathers different phenomena.

This diversity creates an immediate difficulty. If we seek a definition broad enough to include everything, it risks explaining nothing. If we choose a definition that is too narrow - for example "a thought is an inner sentence" - we exclude visual reasoning, gestures prepared without verbalization, or intuitions that precede their formulation.

The aim of this chapter is therefore not to solve the philosophical problem of thought. It is more modest and more useful for our inquiry: to build distinctions that prevent us from confusing a mental phenomenon, its expression, and the model computed from that expression.

We will use the word **thought** to designate a mental activity or state attributed to a person. This definition remains deliberately general. It does not settle the nature of consciousness, nor how a subjective experience emerges, nor the exact boundary between perception, emotion, and reasoning. These questions matter; they go beyond what the evidence mobilized here can establish.

■ UNKNOWN

No brief definition of thought by itself solves the problem of consciousness. The book adopts an operational definition in order to study the relations among mental activities, observable expressions, and models.

Three planes not to confuse

Take an ordinary situation. A person hesitates between two routes, looks at the sky, then says: "I will take the train." We can distinguish at least three planes.

The first is that of **mental experience and activity**. The person may remember a delay, imagine rain, estimate a duration, or feel a preference. Part of this activity may be accessible to them; another may not be clearly so.

The second is that of **observable expression**. The spoken sentence, the glance at the sky, the ticket bought, or the route chosen are accessible to others. They support inferences: the rain may have counted, travel time as well. But several different stories can produce the same choice.

The third is that of **description or model**. An observer may summarize: "this person avoids meteorological risk." A system may estimate a probability of choosing the train. In both cases, a representation is built from cues. It can be useful without being exhaustive.

These planes answer one another, but they are not identical. An expression is produced by a person in a situation; it is not a complete copy of what happens within them. A model organizes observations; it is not the person observed.



Mental activity incomplete strokes: not observed directly



Expression solid stroke: speech, gesture, writing



Model gold rule: constructed description

Figure. Three related but distinct planes: mental activity is not observed directly, expression can be, and the model is built afterward. Author's diagram.

A running example: correcting an equation

Imagine a working page. An equation appears on it. One line is crossed out. Another formulation appears beside it. No letter explains the gesture. Three readings remain possible, and they do not have the same status.

Observation. An equation was written, then corrected on the support.

Inference. The author modified their formulation. Several causes remain compatible with this page alone: error detected, simplification, anticipation of an objection, or simply lack of space.

Hypothesis. This correction could testify to a change in how they treat the problem. This hypothesis is not false in principle. Neither is it established by the crossing-out alone. Surrounding pieces would be needed: a dated version, a letter, a later publication.

Unknown. The exact reason for the correction, the subjective state at the moment of the gesture, and what was never written remain beyond the reach of the available pieces.

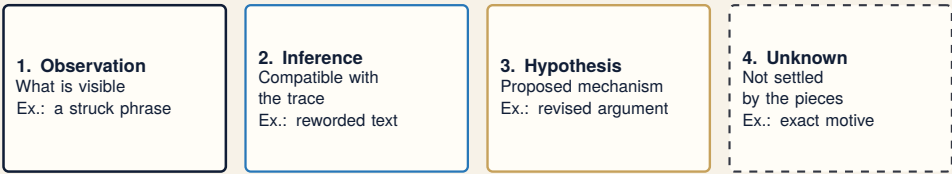


Figure. Four levels for reading a trace: observation, inference, hypothesis, unknown. This methodological principle runs through the book. Author’s diagram.

This grid - observation, inference, hypothesis, unknown - will accompany the rest of the book. It will apply to Einstein’s archives, to an assistant’s replies, and to claims we will be tempted to read as proofs.

The central question is therefore not: “what is thought in general?” It becomes operational: **what can an AI reasonably learn from the traces left by a person, without confusing those traces with the person themselves?**

Expressed thought, attributed thought

We do not wait for advanced technology to attribute thoughts. We do it continually. A friend looks for her keys: we suppose she wants to go out. A colleague rereads a sentence three times: we infer that he doubts his formulation. These attributions make social coordination possible.

They can also be false. The person may have been looking for a document placed near the keys. The colleague may have been correcting a typo without questioning the idea. An observation often underdetermines the explanation: several possible states are compatible with the same trace.

This underdetermination does not condemn every attempt to know another person. It does, however, impose a method. An attribution gains credibility when it rests on independent cues, resists competing explanations, predicts new observations, and can be revised. It remains nonetheless an inference, not unmediated access. Research on mentalizing studies precisely this capacity to represent or infer others’ mental states from cues and from a point of view [SRC-022]. This idea has an older history in work

on “theory of mind,” defined as the attribution of mental states that are not directly observable in order to explain or predict behavior [\[SRC-027\]](#).

Language adds a special case: a person can describe what they think. This testimony is valuable. It can clarify an intention invisible in behavior. But it too is produced in a situation, with a vocabulary, an addressee, and a purpose. We can forget, simplify, reorganize, or discover after the fact an explanation that seems coherent to us. Taking a subjective account seriously does not mean treating it as a perfect and complete transcription.

Representing is not photographing

The term **mental representation** can suggest a small image filed in the head. This metaphor sometimes suits intuition, but it quickly becomes misleading. A representation can be a structure that makes it possible to recognize a situation, compare possibilities, or guide an action without visually resembling what it represents.

The cognitive sciences employ several frameworks to study these structures. The editorial presentation of Johnson-Laird’s book describes a theory in which a person would build inner replicas of the relations among objects and events, then reason from those models. [\[SRC-018\]](#) This formulation comes from the publisher’s notice: the book was not read in full for this audit. Other work distinguishes several memory systems or studies the construction of autobiographical memories. [\[SRC-016\]](#) [\[SRC-017\]](#) These approaches do not reduce to a single theory of thought. They supply testable concepts for particular phenomena.

The word **representation** must therefore always call for two questions: useful for what, and evidenced how? An internal structure can be called a representation if it varies systematically with certain information and contributes to a behavior or a prediction. That finding does not yet say whether it is conscious, lived, or understood as it would be by a human.

This caution will be decisive when we speak of neural networks. The fact that a model encodes regularities is not enough to conclude that it thinks. Conversely, refusing any use of the word “representation” because the system is not human would prevent study of what its internal states actually make possible.

Memory is not an inner archive

We readily imagine a memory as a document preserved and then reopened. The analogy is practical: it gives a beginning, a content, and a place. It can, however, mask the reconstructive character of autobiograph-

ical recall. Work on self-memory describes remembering as the result of an interaction among autobiographical knowledge, present goals, and retrieval cues, rather than as the unchanged reading of a recording. [SRC-017] The psychology of memory also documents several forms of forgetting, distortion, and misattribution. These limits do not make memory useless, but they forbid treating it as a copy [SRC-021].

This idea has two consequences for our project. First, the accounts a person produces about themselves are important data, but situated ones. They can vary without any version necessarily being a lie. Second, a model trained on these accounts learns the regularities of the available accounts. It does not thereby obtain access to an original mental event that would exist intact behind them. An AI can therefore learn forms of storytelling without learning the experience that made them necessary.

These accounts are not only reports on the past. Research on narrative identity also describes them as an evolving way of articulating a reconstructed past and an imagined future [SRC-026]. A model that learns these accounts can therefore capture forms of coherence-making without possessing the lived history that gives them meaning.

The difference between human memory and computer memory must also remain visible. In a digital product, "memory" can designate persistent information recorded and then reinjected into a context. In a person, the word points to multiple biological and cognitive processes. Using the same term proves no identity of mechanism.

What consciousness need not settle here

This book need not settle debates about consciousness in order to proceed. One rule suffices: do not treat an observable performance as automatic proof of a subjective experience. A system can classify, predict, or generate. These capacities are measurable. Whether they are accompanied by an experience comparable to ours is a distinct question, and one unresolved for the models studied here.

■ ESTABLISHED

A behavioral performance and an attribution of subjective experience are two distinct propositions. The first can be measured by a protocol; the second requires additional criteria and remains controversial when it concerns artificial systems.

A working definition

We can now formulate the definition that will guide the chapters that follow:

In this book, a thought designates a mental activity or state attributed to a person. We observe directly only certain expressions and certain effects. From them, we build revisable descriptions and models.

This formulation does not deny the access a person has to their own experience. It clarifies the inquiry's point of view: that of an observer, a historian, a close person, or a system that works from traces.

It also avoids two symmetrical excesses. The first would be to say that, since we do not access another's mind directly, no knowledge is possible. Our relationships, psychology, and history show on the contrary that inferences can be robust. The second would be to believe that a large number of traces abolishes all uncertainty. Quantity sometimes improves estimation; it does not turn observations into identity.

The next chapter turns to the decisive passage between the first and the second plane: how does a human activity become a trace that others can preserve, classify, and interpret?

KEY POINTS

- 01 The book adopts an operational definition of thought, useful for the inquiry.
 - 02 Observation, inference, hypothesis, and unknown are not to be confused.
 - 03 Mental activity, expression, and model are three distinct planes.
 - 04 An AI can learn regularities of traces without accessing the person.
 - 05 Consciousness is not settled by an observable performance.
-

Sources used

- **SRC-016 - Tulving, episodic and semantic memory.** Source to be used only according to the level of examination recorded in the register.
- **SRC-017 - Conway and Pleydell-Pearce, self-memory system.** Used to frame the constructed character of autobiographical memory.
- **SRC-018 - Johnson-Laird, mental models.** Editorial presentation con-

sulted; the full work was not read. The chapter's sentence follows that presentation, not a demonstration verified page by page.

- **SRC-019 - Mitchell and Krakauer, debate on language-model understanding.** Notice and abstract verified; used to distinguish capacities and possible forms of understanding.
- **SRC-021 - Schacter, memory errors.** Notice and abstract examined; used to frame the fallible and constructive character of recall.
- **SRC-022 - Frith and Frith, mentalizing.** Metadata, abstract, and excerpts examined; used to define inference about others' mental states.
- **SRC-026 - McAdams and McLean, narrative identity.** DOI notice and publisher abstract examined; used to link reconstructed past, imagined future, and self-narrative.
- **SRC-027 - Premack and Woodruff, theory of mind.** Notice and Cambridge abstract examined; historical study of a chimpanzee, not direct reading of the human mind.

2 · The mind leaves traces

A crossing-out, a letter, or an equation can document an intellectual path - provided one does not take them for the path itself.

What reaches us

A thought does not arrive in the archives. What arrives there is a re-copied sentence, a crossed-out equation, a letter sent, sometimes a witness's account. Between mental activity and the document, several transformations have already taken place: choosing an expression, inscribing it on a support, preserving that support, classifying it, then making it accessible.

These traces matter. A correction reveals that a formulation was abandoned. Two versions make it possible to follow an evolution. Correspondence places a decision back into an exchange. But no single piece contains, by itself, the totality of the experience that produced it. Silence, hesitation, or an intuition never formulated may leave no mark.

Einstein's archives make this difference concrete. The Einstein Papers Project edits writings, letters, and documents linked to his life and work [\[SRC-001\]](#) [\[SRC-032\]](#). The official notice for volume 5 alone describes more than five hundred letters for the years 1902 to 1914 [\[SRC-036\]](#). That for volume 4 signals notes and calculations that document a research progression toward a relativistic theory of gravitation [\[SRC-035\]](#). This abundance makes it possible to study chronologies, formulations, and exchanges. It does not turn the archive into a continuous recording of Einstein's thought.

This chapter answers a precise question, distinct from that of the next: **what remains of a human activity in the traces preserved?** Documentary transformation is the main object here. Inference about the author will come in chapter 3.

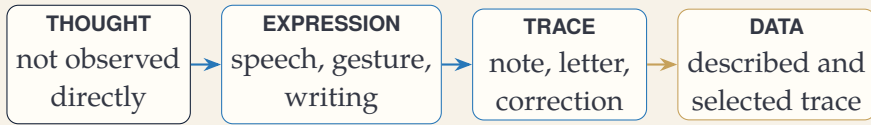


Figure. An expressed thought may leave a trace; a described and selected trace may become data. Each step transforms what will later be interpreted. Author's conceptual diagram.



Each arrow aids study and can also lose context, nuance, or provenance.

Figure. From expression to a usable corpus. Each transformation aids study and may alter or lose information. Author's conceptual diagram.

Expressing already transforms

Moving from an idea to a sentence is not a neutral operation. One must choose a vocabulary, an order, a level of detail, and an addressee. A note to oneself does not resemble an article. A letter to a colleague does not have the same function as a public lecture. The same person can present the same problem in several ways without any of them being the definitive version of their thought.

This transformation is especially visible in drafts. A crossing-out shows a change on the page, but it does not automatically give the reason. Did the author discover an error, prefer a clearer formulation, anticipate an objection, or simply run out of space? The document narrows the set of possible explanations; it does not always choose among them.

Context can help. A letter written the same day, a later version, or a correspondent's reply add cues. Interpretation then becomes a work of relating. It does not consist in making an isolated trace speak, but in comparing pieces whose provenance and date are known.

This rule also holds for digital traces. A very short message can be an urgent request, a reminder, or the continuation of an oral conversation. A history of modifications shows what changed, not necessarily why. The easier a trace is to collect, the more we must resist the idea that it is self-explanatory.

To preserve is to select

We speak of archives as if they passively received the past. In reality, all preservation depends on decisions and circumstances. Some documents are thrown away, lost, or destroyed. Others are kept because they seem important, because an institution acquires them, or simply because their support endures.

Selection continues after preservation. One must inventory, describe, digitize, and sometimes translate. A scholarly edition chooses an order, establishes a text, links variants, and adds annotations. This work greatly increases the corpus's usefulness. It also adds an editorial layer that must remain visible.

In Einstein's case, the *Collected Papers* are not a heap of anonymous files. They constitute a scholarly edition produced by an institutional project, with volumes, commentaries, and links to archival documents [SRC-001] [SRC-033]. The difference is fundamental for an artificial-intelligence system: a passage accompanied by its date, its author, its language, and its documentary status is more interpretable than a string of characters detached from its provenance.

■ ESTABLISHED

An edited archive is a constructed corpus: preservation, description, order, and annotation result from documentary choices. These choices improve access while delimiting what can be observed.

When a trace becomes a datum

A trace is not automatically a datum ready for learning. To be processed by a system, it must be represented in an exploitable form. A page may be photographed, its text transcribed, its language identified, its date encoded, and its relations to other documents described.

Each operation renders possibilities and introduces errors. A handwritten page may pass through physical preservation, digitization, optical recognition, transcription, normalization, indexing, then integration into a corpus. A system that ignores these uncertainties receives an artificially clean version of the material.

Stage	Gain	Risk
Digitization	Access and reproduction	Loss of material detail
Optical recognition	Searchable text	Character confusion
Transcription	Usable text	Normalization or omission
Translation	Linguistic access	Loss of nuance
Metadata	Context and filing	Wrong attribution or dating
Indexing	Fast search	Misrepresented relations

Table. Analysis grid for documentary transformations. Each step may improve access while introducing ambiguity.

Quality therefore does not reduce to the number of documents. It depends on provenance, granularity, metadata, rights of use, and the ability to recover the original support. A million fragments without context can be less useful for certain questions than a smaller set that is dated and annotated. What the assistant in chapter 5 will be able to “learn” will depend directly on these transformations.

One must also distinguish training data from the document consulted at the moment of a reply. In the first case, examples take part in a procedure that adjusts parameters. In the second, an engine retrieves a document and places it in the context. The source text may be identical; its technical role is not.

Negative traces

An archive shows what is present. To reason correctly, one must also consider what is missing. The absence of a letter may mean that it was never written, that it was lost, that it has not yet been found, or that it is not accessible. These hypotheses do not have the same consequences. A retrospective testimony does not automatically fill that void: human recall is itself exposed to forgetting and misattribution [SRC-021].

The same caution applies to regularities. If a person never addresses a topic in the corpus, we cannot conclude that they never thought about it. If they always use a word in public texts, we do not know whether they use it the same way in private notes. A corpus describes a distribution of traces, not the set of all possible situations.

These absences can be systematic. Activities regarded as ordinary sometimes leave fewer documents than public events. People with institutional means are better archived. Certain languages, supports, and groups survive more in collections. A model built on archives inherits this uneven geography.

Recognizing these limits does not condemn the project. It makes it possible to ask more precise questions: which periods are covered? Which

genres of documents dominate? Which intermediaries selected the pieces? Which silences are known, and which remain invisible?

Reading a path without inventing its steps

Suppose three versions of a line of reasoning are preserved. The first contains an analogy, the second an additional equation, the third removes the analogy. We can establish a documentary sequence if dates or material cues allow it. We can describe the differences. We can sometimes link a change to a letter or to a known experiment.

We must be more cautious when we narrate what happened "in the author's head." Saying that they removed the analogy is an observation. Saying that they judged it misleading is an inference. Saying that a precise intuition appeared to them at that moment may be a hypothesis without evidence.

The book's epistemic vocabulary serves exactly this purpose. It does not forbid narrative; it indicates its solidity. A history of ideas becomes more interesting, not less, when the reader knows where the document ends and where interpretation begins.

What an assistant could exploit

An assistant linked to a well-documented corpus could accomplish several useful tasks. It could retrieve all occurrences of a term, bring versions together, build a provisional chronology, or flag an apparent contradiction. It could propose relevant passages to a historian, who would return to the documents.

It could also generate an erroneous synthesis, merge two periods, or present an inference as a fact. The fluency of the text does not guarantee archival fidelity. For this reason, every important reply should keep a path to the pieces used and distinguish citation, paraphrase, and interpretation.

A personalized model could learn more abstract regularities: sentence length, order of arguments, types of examples, or ways of answering an objection. These regularities belong to the productions observed. Calling them a "style of thought" can be useful in a limited sense, provided one does not confuse them with the totality of thought.

■ INFERENCE

Numerous, dated, and diversified traces can support a more robust model of certain intellectual regularities. They guarantee neither the exhaustiveness of the corpus nor access to the subjective experience that accompanied their production.

From trace to model

The decisive passage can now be formulated. An expressed thought sometimes produces a trace. That trace can be preserved and described. A selection of traces can become a corpus. A system can finally build, from that corpus, a representation that predicts or generates certain productions.

At each step, something is gained: stability, comparability, search, computation. At each step, something can be lost: context, ambiguity, silence, provenance, or singularity. Rigor consists in documenting these transformations rather than erasing them under the word "data."

The next chapter will start from the corpus thus constituted. Even if the traces are authentic and carefully edited, how far do they allow us to know a person? How does one distinguish a useful model from a circular portrait that only recovers its designers' expectations?

KEY POINTS

- 01 This chapter studies documentary transformation, not yet inference about the author.
 - 02 Each step of access can also lose, modify, or render ambiguous.
 - 03 Provenance and metadata matter as much as volume.
 - 04 The absence of a trace is not proof of the absence of a thought.
 - 05 What a system will be able to exploit depends on these transformations.
-

Sources used

- **SRC-001 - Einstein Papers Project and Einstein Database.** Institutional presentations consulted; they support the corpus's scope, its scholarly organization, and the general categories of documents.
- **SRC-021 - Schacter, memory errors.** Notice and abstract examined; used only to recall the limits of retrospective testimony.

- **SRC-004 - Lewis et al., retrieval-augmented generation.** Abstract verified; used only to distinguish documentary retrieval and parametric memory.
- **SRC-032 - Einstein Papers Project, history of the project.** Institutional page consulted in full; it documents the history of organizing and editing the papers.
- **SRC-033 - Collected Papers, volume 1.** Official editorial notice consulted; it characterizes the scope of the early correspondence without attributing unverified content.
- **SRC-035 - Collected Papers, volume 4.** Official editorial notice consulted; it supports the presence of research notes and calculations for 1912-1914.
- **SRC-036 - Collected Papers, volume 5.** Official editorial notice consulted; it supports the general scope of more than 500 letters, not their individual content.
- **Complementary archival sources.** To be integrated after verification of the phase-3 research batch; no precise historical citation is attributed without a located document.

3 · Can one know someone without being them?

A model can become precise, useful, and surprising without ceasing to be an inference built from a limited point of view.

After the trace, the inference

The previous chapter followed the document: expression, preservation, digitization, metadata, corpus. This one changes its object. The question is no longer: *what remains?* It becomes: **what can we deduce from these traces about their author?**

We know people we are not. We anticipate a friend's reaction, recognize a colleague's style, and understand that an ironic sentence should not be taken literally. This knowledge is neither perfect nor imaginary. It rests on a history of interactions, words, gestures, and shared situations.

It remains, however, indirect. We do not feel another's pain in their place. We do not see all their memories at the moment they decide. We build hypotheses about their intentions and beliefs, then revise them when their words or actions contradict them. This inference about others' mental states is studied under the name of mentalizing [\[SRC-022\]](#). The strategy of explaining or predicting behavior by attributing beliefs and desires also has a philosophical history [\[SRC-024\]](#).

The question is therefore not to choose between two extremes - knowing perfectly or knowing nothing. It is to determine what a model allows us to foresee, in which situations, with what uncertainty, and at the cost of which simplifications.

This question precedes artificial intelligence. A biographer, a physician, a close person, and a recruiter all build partial representations of a person. Their data, goals, and responsibilities differ. A computed model makes this construction more explicit and sometimes easier to deploy at scale. It does not make it automatically neutral.

From portrait to predictive model

A portrait often seeks to make a person intelligible. It selects events, traits, and tensions. A predictive model pursues a more operational goal: from past observations, to estimate a reply, a choice, or a preference in a defined situation.

The difference matters. A text can be a convincing portrait without allowing one to predict a new reaction. Conversely, a model can correctly predict limited choices without telling a faithful story of the person. It can exploit statistical regularities whose human interpretation remains uncertain.

Studies of digital traces show the power and the narrowness of this kind of inference. In defined datasets, Facebook "Likes" have made it possible to predict statistically certain attributes or psychometric traits [\[SRC-028\]](#) [\[SRC-029\]](#). These results concern specific targets, populations, and a platform. They demonstrate neither a global understanding of the person nor the reconstruction of their lived history.

Work gives an experimental form to this idea without resolving the human case. Rabinowitz and colleagues trained a network to predict artificial agents observed in simple grid worlds [\[SRC-023\]](#). Park and colleagues built an environment in which twenty-five generative agents combined a natural-language history, memory retrieval, a reflection step, and planning [\[SRC-012\]](#). These demonstrations concern delimited tasks and architectures. They prove neither the reconstruction of a real person nor access to their experience. Language models have also reproduced certain distributions of group responses or certain experimental results [\[SRC-030\]](#) [\[SRC-031\]](#). Again, simulating a distribution or an average response is not enough to establish fidelity to a particular individual.

To model certain intellectual productions, we could gather texts, corrections, replies to objections, and bibliographic choices. The system would seek regularities: associated concepts, order of arguments, preferred examples, degree of caution, or ways of revising a conclusion.

The result would not be "the person." It would be a function built for certain tasks from a determined corpus. This formulation may seem cold. It nonetheless protects against a major error: attributing to the model everything we know, love, or imagine about the individual.

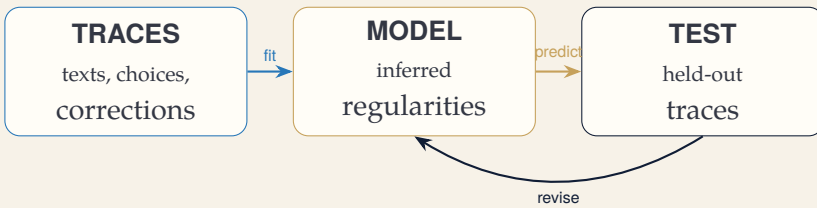


Figure. A model of a person must be evaluated on productions not used in its construction. Resemblance is not proof of identity. Author's conceptual diagram.

Three documentary situations

Comparing three cases makes inference measurable.

An isolated document. A sentence appears in a letter. Several explanations remain open. The same sentence, read alone, can be a polite formula, a scientific position, or a tactical concession.

Several independent documents. A letter, a calculation note, and a public text converge. The hypothesis gains strength if the pieces do not repeat the same source in different forms. A hundred copies of the same account are not a hundred proofs.

Contradictory documents or documents from different periods. Two dated formulations oppose each other. The inference must then name the period, expose the tension, or abstain. A seamless synthesis is not a resolution: it is a loss of information.



Figure. An inference grows firmer when documents are contextualized, cross-checked, and tested on new items. It still does not become certainty. Author's diagram.

The genre of the document matters as much as its content. A sentence read in a private letter, a draft, or a public text does not authorize the same

conclusions about the author. Attributing a precise psychological intention to Einstein from a single volume notice would here be a methodological fault.

How to test a model of a person

The weakest test consists in asking the model to reproduce what it has already seen. A system could restore a formula or imitate a passage without having learned a generalizable regularity. To evaluate further, one must reserve productions that were not used in its construction.

Suppose we have dated letters. One part forms the working corpus; another remains hidden during development. The model receives the context of a held-out letter and must predict, not every exact word, but certain elements defined in advance: position defended, type of argument, concept mobilized, or degree of confidence.

Criteria must be fixed before looking at the result. Otherwise, a vague reply can always be declared "resembling" after the fact. One must also compare the system to baselines: a simple rule, non-specialist readers, experts familiar with the corpus. A result is informative only relative to a comparison base.

Finally, evaluation must seek failures. What happens when the context changes, when a question is ambiguous, or when two periods of life present different positions? A model that mixes these periods can seem coherent while being historically false.

■ ESTABLISHED

Evaluating on held-out data reduces the risk of confusing restitution and generalization. This principle is not enough: criteria, comparisons, and error analysis determine what the test actually allows one to conclude.

The person changes

A model of a person often contains a silent hypothesis: the individual would be stable enough for the past to predict the future. This hypothesis is sometimes reasonable. A lasting preference or an acquired skill creates regularities. It becomes misleading if it erases learning, ruptures, and changes of context.

A researcher's writings early in their career do not necessarily describe their later positions. A disagreement can modify their method. New evidence can reverse a conclusion. Even vocabulary evolves. A model trained

on all periods without chronology risks producing an average that never existed.

Temporality must therefore be part of the data. One must be able to ask: “according to the documents of which period?” A reply that acknowledges evolution is more faithful than an artificially uniform voice.

This problem also exists for contemporary personalized assistants. A preference recorded yesterday can become obsolete. Accurate information can cease to be so. The ability to consult, correct, and delete a memory is not only a practical function; it answers the changing nature of the person represented.

The evaluators’ mirror

To judge whether a model “resembles” someone, one needs evaluators. They bring their own expectations. A public may recognize above all the most famous traits of a personality. A specialist may favor a period or a historiographical interpretation. A close person may value aspects absent from public archives.

The model then risks becoming a mirror of stereotypes. If it produces a sentence associated with Einstein’s popular image, it may seem right even if it corresponds to no attested document or reasoning. The more famous the figure, the more intuitive recognition must be completed by documentary criteria.

A robust evaluation therefore separates several questions:

- is the text compatible with the sources of the period?
- does it reproduce regularities that were not explicitly provided?
- does it cite provenance correctly when it asserts a fact?
- does it know how to signal that a reply is not determined by the corpus?
- do experts agree, and on which criteria?

Disagreement among evaluators is not only noise to eliminate. It can reveal that the notion of fidelity has not been sufficiently defined.

Personality, identity, role

Another shortcut consists in speaking of an “intellectual twin.” The term evokes a global copy. In practice, a system will generally be better at certain roles: retrieving passages, simulating a form of pedagogical dialogue, proposing an objection compatible with a corpus, or imitating a style.

These roles do not constitute an identity. A person's identity includes a lived history, relationships, a body, commitments, and a capacity to answer for their acts. A system that generates sentences does not become the historical subject whom those sentences resemble.

The distinction has ethical consequences. Presenting a simulation as a deceased person can create a misleading authority. The user may attribute to the individual remarks never made. Historical disagreements may be smoothed to produce a more agreeable voice. An interface must therefore announce the system's nature, its sources, its period of coverage, and its limits. The vocabulary chosen matters too: speaking as if the system were spontaneously a person can make one forget what its mechanisms and evaluations actually establish [\[SRC-025\]](#).

■ **ESTABLISHED**

A similarity of style or reply does not establish identity between a model and a person. The status of simulation, the corpus used, and the provenance of claims must remain visible.

What a model can know

We can now give a conditional answer to the chapter's question. Yes, a model can capture real regularities in a person's productions. It can sometimes anticipate a choice, recover a recurrent structure, or propose a reply that evaluators judge compatible with a corpus.

No, that success does not give it total access to the person. The model is limited by the available traces, the way they were selected, the situations covered, and the training objective. It can be precise and yet silent about what was never expressed. It can be convincing and yet false outside its test distribution.

The right question is therefore not: "does the system really know this person?" posed without further precision. One must ask:

Which characteristics of the productions can it estimate, with what reliability, under which conditions, and which observations could show that it is wrong?

This formulation turns an impression into a research program.

The threshold of the thought experiment

Return to Einstein. We have a rich but incomplete historical corpus, edited by institutions and crossed by several languages, periods, and gen-

res [SRC-001] [SRC-032] [SRC-036]. We could imagine a system that explores this corpus and generates replies. To know what it represents, one would need to separate retrieved documents, learned regularities, and evaluators' expectations.

Above all, we should preserve the boundary between two Einsteins. The historical Einstein appears in attributed and contextualized documents. The imagined Einstein serves to pose contemporary questions. A generated reply belongs to neither: it is the output of a system built by us.

This boundary opens the next part. We can now examine the thought experiment in detail: which assistant would we build, which traces would we give it, and how would we avoid taking its voice for a resurrection?

KEY POINTS

- 01 After the transformation of traces comes inference about the author.
- 02 An isolated document, independent pieces, and contradictions do not have the same power.
- 03 Multiplying copies of the same source is not equivalent to multiplying proofs.
- 04 A model of a person must announce task, period, and corpus.
- 05 The similarity of an output establishes neither identity nor historical authority.

Sources used

- **SRC-001 - Einstein Papers Project.** Used to characterize the historical corpus without attributing a fictional reaction to Einstein.
- **SRC-012 - Park et al., Generative Agents.** Metadata, abstract, and method excerpts examined; example of agent architecture, not a faithful model of a real person.
- **SRC-017 - Conway and Pleydell-Pearce.** Mobilized for the relation among autobiographical memory, present goals, and representation of the self.
- **SRC-022 - Frith and Frith, mentalizing.** Metadata, abstract, and excerpts examined; frame for inference about others' mental states.

- **SRC-023 - Rabinowitz et al., Machine Theory of Mind.** PMLR notice, abstract, and introduction excerpts examined; example limited to artificial agents in grid worlds.
- **SRC-024 - Dennett, intentional stance.** Notice and excerpts examined; philosophical position, not an experimental result.
- **SRC-025 - Shanahan, language used about LLMs.** Metadata, abstract, and conclusion excerpts examined; used to prevent unjustified anthropomorphism.
- **SRC-032 - Einstein Papers Project, history of the project.** Institutional page consulted in full; it documents editorial history, not every document.
- **SRC-036 - Collected Papers, volume 5.** Official editorial notice consulted; it specifies the general scope of a correspondence corpus.
- **SRC-028 - Youyou, Kosinski and Stillwell, personality judgments.** DOI notice and PubMed abstract examined; results limited to a sample, a platform, and psychometric targets.
- **SRC-029 - Kosinski, Stillwell and Graepel, digital traces.** DOI notice and PubMed abstract examined; probabilistic predictions, not individual certainties.
- **SRC-030 - Argyle et al., simulation of human samples.** Notice and Cambridge abstract examined; collective distributions, not personal fidelity.
- **SRC-031 - Aher, Arriaga and Kalai, replication of human studies.** PMLR notice and abstract examined; the PDF is not claimed as fully read.

4 · Einstein in 2026

The scenario can illuminate a method, provided one never treats the imagined interface as an archive.

■ THOUGHT EXPERIMENT

Einstein in 2026 is a methodological fiction. No gesture, question, or reply described in this part is a historical fact.

Two presences in the same room

Imagine the workspace again. The screen is on. The notebook is open. This time, the question is no longer what a thought is in general. It is which Einstein we would actually have placed before this screen.

There are at least two, and they must not overlap. The first is historical. He died in 1955. We know him through edited, dated, and annotated documents: letters, articles, notes, and calculations [\[SRC-001\]](#) [\[SRC-032\]](#). The second is a method character. He serves to ask what a contemporary assistant could do with these documents. He cannot testify in place of the first.

This separation seems simple. It is nonetheless the condition of the entire part. If we relax it, a fluent reply may be read as a citation, and a real citation may be drowned in an invented scene.

What the archives already allow us to see

The notices of the *Collected Papers* give an idea of the material's diversity, without replacing the pieces themselves. The volume of the early years notes early correspondence and an interest in the electrodynamics of moving bodies [\[SRC-033\]](#). The volume of the Swiss writings gathers scientific work from 1900 to 1909, including the 1905 articles [\[SRC-034\]](#). Volume 4 indicates notes and calculations linked to the progression toward a relativistic theory of gravitation [\[SRC-035\]](#). Volume 5 describes more than five hundred letters for 1902-1914 [\[SRC-036\]](#).

These indications authorize a limited conclusion: the corpus is not a single public voice. It comprises different genres, periods, and addressees. They do not authorize narrating the content of a precise letter or the exact path of a calculation so long as the piece has not been consulted.

■ **ESTABLISHED**

Official notices describe the scope of volumes and documentary categories. They do not replace reading the numbered documents.

Documentary box - the hypothetical corpus

For the thought experiment, the assistant would not have "Einstein." It would have an edited corpus, with distinct categories:

- **Scientific texts** published or prepared in a named period.
- **Correspondence** dated, with variable addressees and languages.
- **Notes and calculations** noted in certain volume notices.
- **Editorial commentaries** that are not Einstein's voice.
- **Translations** and later presentations, to be labeled as such.

The Einstein Papers Project provides an institutional source for verifying the scope of the archives and the characteristics of the volumes [\[SRC-001\]](#) [\[SRC-032\]](#). Access to an edited collection is not equivalent to possessing all the traces of Einstein's intellectual life. Absences, undigitized pieces, and editorial choices remain.

The inquiry's concrete question

Before speaking of a model of thought, the assistant receives a narrow task:

From a set of documents dated in a defined period, can one

describe how a scientific formulation evolves between two versions

without inventing the cause of the change?

This question extends the example from chapter 1. It does not require reconstructing a consciousness. It requires distinguishing documentary observation, inference, and unknown - then citing the pieces.

Myth is also a trace

An assistant fed only on famous anecdotes would not necessarily learn the documented method. John D. Norton argues, in the introduction to a 2016 article, that the discoveries of 1905 were preceded by long inquiries, often ordinary in their labor, and that several popular narratives reduce them to shortcuts [\[SRC-040\]](#). He announces four myths: childish thought, purely operational thought, the mere pleasure of breaking rules, and the power of a unique experiment.

This thesis is that of the article, established from its abstract and introduction. The case studies that follow have not been examined in full here. It is nonetheless enough to pose a requirement: a model of Einstein should be able to be confronted with working documents, not only with the image of instantaneous genius.

The risk is concrete. If the most available data are simplified biographies, posters, or detached sentences, the system will first reproduce that cultural circulation. It may seem convincing precisely because it confirms an already widespread expectation.

What the 2026 scene adds

In the fiction, the imagined Einstein could question the assistant, correct a reply, or leave a question unanswered. These gestures would create new traces. They would not belong to the historical archives. They would belong to the scenario and to the system that answers in it.

One would therefore need to keep two journals. One records historical documents and their provenance. The other records imagined or simulated interactions, with the date of the simulation, the model used, and the texts supplied in the context. Mixing these journals would produce a fictional archive.

A thought experiment remains useful because it organizes a scenario in order to examine consequences [\[SRC-038\]](#). Here, the consequence to examine is practical: which categories of documents, which metadata, and which tests would be necessary before speaking of an intellectual assistant?

What would remain unknown

Even a very rich corpus would leave closed zones. We would not know which hesitations were never written. We would not know which formulation was chosen for an addressee rather than because it seemed the most exact. We would not know how Einstein would have reacted to an interface that did not exist.

These unknowns do not make the project vain. They fix its edge. The next chapter can therefore leave the character and describe the tool: not what Einstein would have said, but which distinct mechanisms an assistant should expose for its replies to remain interpretable.

KEY POINTS

- 01 Einstein in 2026 is a fictional device, distinct from the documented Einstein.
 - 02 The hypothetical corpus must separate genres, periods, and editorial commentaries.
 - 03 The concrete question concerns the documentary evolution of a formulation, not a resurrection.
 - 04 Myths of discovery can become more visible data than the dossiers.
 - 05 The next chapter describes the tool's architecture.
-

Sources used

- **SRC-001 - Einstein Papers Project.** Institutional scope consulted; it is not an inventory of every document.
- **SRC-032 - History of the project.** Institutional page consulted in full.
- **SRC-033 - Volume 1.** Official editorial notice; individual documents not consulted.
- **SRC-034 - Volume 2.** Official editorial notice; the commercial date may differ from the edition date and must remain flagged.
- **SRC-035 - Volume 4.** Official editorial notice; manuscripts not surveyed page by page.
- **SRC-036 - Volume 5.** Official editorial notice; no individual letter is cited from this notice alone.

- **SRC-038 - Thought experiments.** Academic entry consulted in full to frame the scenario's status.
- **SRC-040 - Norton, myths of discovery.** Abstract and introduction examined; later case studies are not claimed as fully read.

5 · Einstein's personal chatbot

An intellectual assistant is not a person reduced to a window: it is an assembly of corpus, context, memory, and search.

■ THOUGHT EXPERIMENT

The chatbot described here has not been built. The chapter specifies what would need to be separated for such a construction to remain intelligible.

One window, four operations

The previous chapter fixed the character. This one fixes the tool. In the fictional scene, a single window would receive the question. That interface simplicity must not decide the architecture.

Four operations remain distinct. The **parameters** come from prior training. The **context** gathers what is supplied for the reply under way. A **product memory** can keep certain information and reinject it later, according to settings. An **external search** can retrieve documents and place them in that context.

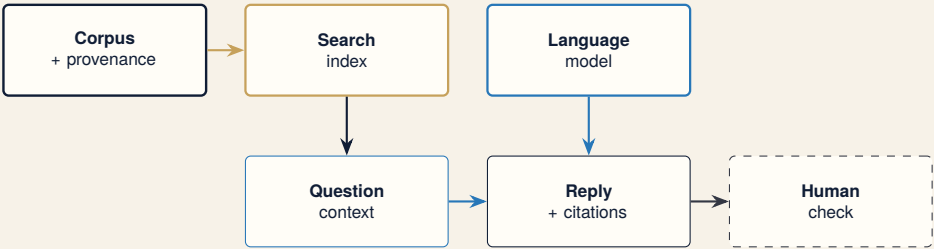
ChatGPT's documentation, consulted on October 9, 2026, separates memory controls from controls concerning the use of conversations to improve models [\[SRC-006\]](#) [\[SRC-007\]](#). This separation is a product policy, dated and changeable. It is not a law of all systems.

■ ESTABLISHED

In the documents consulted, memorizing a preference, retrieving a document, and adjusting weights are not the same operation.

Hypothetical functional architecture

To make the inquiry visible, the imagined assistant can be described as a chain of functions, not as a person in a window.



Proposed architecture for a thought experiment. Neither ChatGPT nor a real historical system.

Figure. Hypothetical functional architecture of a historical assistant. Proposal for the thought experiment, not a description of an existing product. Author’s diagram.

Function	Question	Limit
Search	Where is the relevant document?	The passage may be inadequate
Contextualization	Which period or situation?	Context may be incomplete
Comparison	What differs across items?	Cause is not automatic
Generation	How to phrase a reply?	A sentence may outrun the evidence
Verification	Do sources support the claim?	Verification itself must be checked

Table. Functions of a hypothetical historical assistant. Proposed architecture, not a system actually built for Einstein.

This architecture is a proposal for the thought experiment. It does not identify ChatGPT. It does not describe a historical system actually built for Einstein. It serves to show what would need to be exposed for a reply to remain interpretable: corpus, search, context, generation, citations, verification.

Adjusting weights is not conversing

The difference becomes clearer if one looks at a training job. OpenAI’s supervised fine-tuning documentation describes a sequence of steps: assemble examples, upload them, create a training job, then evaluate the resulting model. It presents that job as what updates the weights from the examples [SRC-039]. The same page indicates that the availability of this platform is shrinking. One must therefore date the information and not turn it into a permanent description of ChatGPT.

The result from Brown and colleagues points in the same direction for another situation: the few-shot tasks evaluated in their article are performed without a gradient update during those tasks [SRC-009]. An

example placed in a prompt can change the present reply. It does not by that alone constitute a new training of the general model.

A personal assistant could use both regimes, but it should say so. "This reply rests on retrieved passages" does not mean the same thing as "the weights were adjusted on a historical corpus."

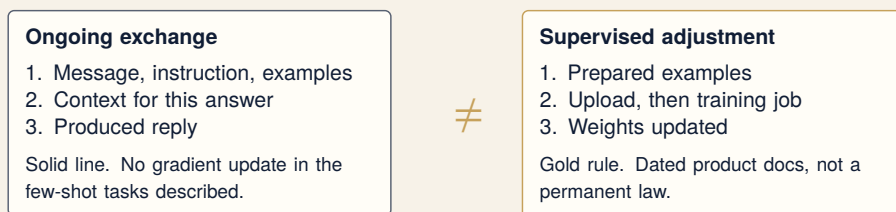


Figure. Left: the message changes the context of the reply. Right: a training run updates weights. Explanatory diagram. [SRC-009], notice and summary. [SRC-039], page consulted on 9 October 2026.

The corpus should be described before it is queried

Before any question, there should be a corpus sheet. It would indicate the period, the languages, the genres, and the status of each piece: published article, letter, calculation note, translation, or editorial commentary. It would also distinguish Einstein's text from introductions written by the editors.

The notices already consulted show that these genres exist and that they are not interchangeable [SRC-033] [SRC-034] [SRC-035] [SRC-036]. A letter to a close person and a scientific memoir do not supply the same type of cue. A calculation note can document an attempt. A public text can simplify, for a reader, a result long prepared [SRC-040].

The minimal metadata would therefore be the date, the attributed author, the addressee when there is one, the language, the genre, the edition used, and the state of verification. Without them, a search could bring together sentences that nothing, historically, brings together.

Personalizing is not resurrecting

A product memory can retain that a user prefers short replies or always wants references. This function is useful. It produces a usage profile. It does not produce the autobiographical memory of the person whose archives one studies.

The generative agents of Park and colleagues combine a natural-language history, retrieval, reflection, and planning. Their demonstration

concerns twenty-five agents in a controlled environment [SRC-012]. It shows that an architecture can make a behavior more consistent. It does not show that an agent becomes the person whose certain roles it imitates.

For Einstein, the rule should be visible in the interface itself: corpus covered, period covered, documents retrieved, generated parts, and level of uncertainty. A continuous voice can mask these layers. The next chapter asks how to measure whether, despite this transparency, the system learns something other than a surface imitation.

KEY POINTS

- 01 The hypothetical architecture separates corpus, search, model, context, and verification.
 - 02 Each function has a limit: retrieving is not proving.
 - 03 Updating weights remains distinct from an ongoing exchange.
 - 04 A product's personalization is not a person's lived memory.
 - 05 The next chapter turns this architecture into an evaluation protocol.
-

Sources used

- **SRC-006 - OpenAI data controls.** Official page consulted on October 9, 2026.
- **SRC-007 - OpenAI memory.** Official page consulted on October 9, 2026; functions vary by offering.
- **SRC-009 - Brown et al.** Notice and abstract; no gradient update during the few-shot tasks described.
- **SRC-012 - Park et al.** Abstract and method excerpts; twenty-five fictional agents.
- **SRC-033, SRC-034, SRC-035, and SRC-036 - Volume notices.** Documentary categories, not the content of each piece.
- **SRC-039 - OpenAI supervised fine-tuning.** Page consulted on October 9, 2026; platform availability evolving.
- **SRC-040 - Norton.** Introduction used only to recall that public texts can simplify long work.

6 · The Einstein model

A model has scope only if one can say which task it performs, over which period, and against which held-out pieces.

■ THOUGHT EXPERIMENT

No Einstein model is evaluated in this chapter. The protocol described is a research proposal to be built, not a result obtained. No score is invented.

Name the task before the system

“Modeling Einstein” is a formula too broad to be tested. It mixes style, concepts, changes of mind, relationships, and lived experience. A protocol must choose a narrow task.

Three candidate tasks, mutually exclusive for a first trial:

1. **Documentary genre.** From a dated excerpt, estimate whether it is rather a calculation note, a public text, or a letter, according to the categories already visible in the volume notices [\[SRC-033\]](#) [\[SRC-035\]](#) [\[SRC-036\]](#). 2. **Period coherence.** Say whether a proposed formulation is compatible with the dossier of a named period, without presenting it as a sentence of Einstein’s. 3. **Provenance.** Distinguish a claim documented in the supplied corpus from an invented continuation.

The task must announce its period. The notices already show that 1900-1909 and 1912-1914 do not present the same dossiers [\[SRC-034\]](#) [\[SRC-035\]](#). A system that answers with the voice of one period to a question posed in another can seem coherent while being historically false.

■ HYPOTHESIS

A system built on selected traces could reproduce certain regularities of those traces without reproducing Einstein’s subjective experience. This possibility remains to be tested.

Proposed protocol - eight boxes

Here is the minimal grid. It has not been executed in this book.

1. **Question.** Can one predict a defined documentary property without reconstructing a person? 2. **Corpus.** Identifiable scholarly edition; genres separated; editors' commentaries labeled as such [SRC-001]. 3. **Period.** A single chronological window per experiment. 4. **Separation.** Preparation pieces, development pieces, evaluation pieces held out. 5. **Configurations.** (a) documentary search alone; (b) few-shot context without a gradient update during the task, regime described for GPT-3 [SRC-009]; (c) possible weight adjustment on a set of examples, distinct regime [SRC-039]. 6. **Comparison base.** Human readers with the same dossier, and a simple term-frequency rule. 7. **Criteria fixed before reading.** Accuracy on the task, citation of provenance, stability under reformulation, rate of inventing absent pieces. 8. **Publication of failures.** Change of language, addressee, period, editor / author confusion.

This separation reprises the distinction between retrieving a document and adjusting weights [SRC-004] [SRC-039]. A search system can cite the held-out piece if it has access to it at test time. That is not generalization. Generalization begins when the piece has served neither for the reply index nor for adjustment.

Stage	Requirement	Typical failure
Task	A precise documentary property	"Model Einstein" with no criterion
Period	A single chronological window	Mixing 1905 / 1913
Hold-out	Items kept out of construction	Leakage via search or context
Criteria	Accuracy, provenance, stability	Judging a "right voice"

Table. Minimal grid for the protocol proposed in chapter 6. Research proposal, not a result obtained.

A single main task for the first trial

To avoid diluting the protocol, the first experiment retains **documentary fidelity**: can the system answer a question about a given period by citing the documents that support its claims?

This target is more controllable than "reproducing Einstein's thought." It reprises the question from chapter 4 on the evolution of a formulation. It requires verifiable citations, a named period, and a capacity to abstain when the corpus does not suffice.

Data separation

Three sets must remain distinct before any reading of results.

1. **Construction.** Pieces used to build the documentary index or, where applicable, for a weight adjustment. 2. **Development.** Pieces used to tune prompts or thresholds, without entering the final score. 3. **Evaluation.** Pieces held out. A piece accessible via search at test time cannot be presented as proof of generalization independent of that piece.

Criterion	Question	Typical failure
Accuracy	Correct given the items?	Plausible invention
Provenance	Do the citations exist?	False reference
Temporality	Is the period respected?	Mixed dates
Robustness	Survives reformulation?	Wording success
Abstention	Signals lack of evidence?	Forced answer
Contamination	Leakage of test items?	False generalization

Table. Proposed evaluation grid for the chapter 6 protocol. Criteria to fix before inspecting results. No score was obtained here.

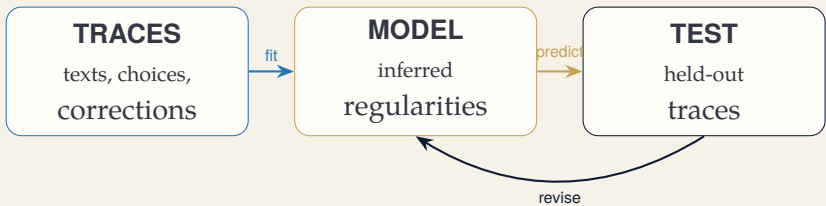


Figure. A model of a person must be evaluated on productions not used in its construction. Resemblance is not proof of identity. Author's conceptual diagram.

Error categories to anticipate before the test

A reply can fail for different reasons. Naming them in advance prevents treating all failures as a single verdict.

- **Wrong document:** citation of a piece off-topic or out of period.
- **Temporal confusion:** mixing of two moments in the corpus.
- **Interpretation presented as fact:** an inference is no longer labeled.
- **Smoothing synthesis:** a documented contradiction disappears.
- **Invented provenance:** elegant reference absent from the archives.

- **Contamination:** the reply rests on a piece reserved for the test.

What would count as success, and as failure

Limited success. On the documentary-fidelity task, the system cites correctly, respects the period, abstains when pieces are missing, and beats a simple frequency rule without inventing letters.

Informative failure. The system invents a provenance, mixes 1905 and 1913, attributes to Einstein an editor's commentary, or succeeds only when the answer is shown in the context.

What would never count as evidence. A voice that "sounds right." An imaginary recognition. An impression of understanding. The next chapter explains why these judgments, even sincere, do not replace the grid above.

The scale given by existing simulations

A few results give a scale, not an Einstein model. Rabinowitz and colleagues predicted artificial agents in grid worlds [SRC-023]. Park and colleagues obtained behaviors judged consistent in fictional agents [SRC-012]. Argyle and colleagues showed that GPT-3, conditioned by profiles, could reproduce certain distributions of responses from American groups [SRC-030]. Aher, Arriaga and Kalai reproduced certain study results and also observed a distortion of over-precision [SRC-031].

None of these works studies Einstein. Their common lesson is narrower: a simulation can be useful for a task and remain unfaithful to a person. It can even be too regular, where human responses disperse.

Trait predictions from digital traces call for the same reservation. In precise Facebook data, certain attributes or personality judgments have been predicted statistically [SRC-028] [SRC-029]. Performance depends on the population, the platform, and the target. It does not reconstitute an intellectual biography.

What this chapter adds

The previous chapters fixed the character and the tool. This one fixes the test bench. Without task, period, hold-out, and failure criteria, "the Einstein model" remains a slogan. With them, it becomes a hypothesis that one can, one day, put to the test.

KEY POINTS

- 01 Task, period, and corpus must be named before evaluation.

02 Test pieces must remain outside construction and the search used to answer.

03 A convincing voice is not a criterion.

04 Existing simulations concern other agents or groups.

05 The protocol in this chapter has not been executed here.

Sources used

- **SRC-001 - Einstein Papers Project.** Scope of an edition, not piece-by-piece reading.
- **SRC-004 - Lewis et al.** Abstract; parametric memory and retrieval.
- **SRC-009 - Brown et al.** Notice and abstract; few-shot without a gradient update during the tasks described.
- **SRC-012, SRC-023, SRC-030, SRC-031.** Limited simulations; none on an Einsteinian corpus.
- **SRC-028 and SRC-029.** Statistical predictions in specific digital data.
- **SRC-033 to SRC-036.** Volume notices for genres and periods.
- **SRC-039 - Supervised fine-tuning.** Page consulted on October 9, 2026.

7 · Would Einstein have recognized his own model?

Imaginary recognition can reveal our expectations. It cannot serve as historical proof.

■ THOUGHT EXPERIMENT

No recognition by Einstein is available as a fact. The question serves to examine the value and the weakness of that criterion.

The judge we would like to have

After a successful test, a question returns almost naturally: would Einstein have recognized this answer as his own? It is powerful because it promises an ultimate witness. It is also unusable as proof. The historical witness cannot compare the system's output to his intention. Making him answer inside the fiction would only create a second generated text.

Recognition, even when it concerns a living person, is not a simple instrument. A person may recognize themselves in a flattering portrait and reject an exact one. They may forget an earlier position or prefer the wording they would give today. For a posthumous corpus, that difficulty becomes a practical impossibility.

■ UNKNOWN

No protocol makes it possible for historical Einstein to recognize a model built after 1955. The question must remain a critical tool, not a test.

Five senses of fidelity

The word "recognize" conceals several judgments. They are not equivalent.

Fidelity	What would be assessed	What it does not prove
Stylistic	Vocabulary and form	Deep historical fidelity
Documentary	Accuracy of citations	Full understanding
Historical	Period and corpus coherence	Lived experience
Scientific	Validity of a demonstration	Einstein's historical method
Behavioral	Regularities across tasks	Personal identity

Table. Five senses of fidelity. Confusing them turns intuitive recognition into proof.

A reply may resemble the expected style without being consistent with the documents. It may cite correctly without restoring a historical path of inquiry. It may solve a scientific problem according to defined criteria without reproducing the path Einstein would actually have followed. The protocol of chapter 6 becomes necessary precisely because these judgments are too easily confused.

Resemblance and stereotype

An output may be recognized by readers for the wrong reasons. It confirms the image they expect: a sudden illumination, a rule lightly broken, a unique experience standing in for inquiry. Norton presents precisely these narratives as simplifications able to eclipse years of documented work [SRC-040]. A model trained on the most repeated versions of those narratives would have a good chance of reproducing them. Its success would then measure the force of the stereotype, not fidelity to the dossiers.

One must therefore ask who recognizes, from which sources, and in competition with which other reply. Historians with held-out pieces do not share the same judgment as readers fed on the public image. Their disagreement is not noise to erase. It shows that the word “resemblance” covers several targets.

Work on simulating groups recalls that a plausible distribution can still be too smooth or too precise [SRC-030] [SRC-031]. Applied to a person, that lesson forbids taking fluency for an admission of identity.

What a disagreement would teach

In the fictional scenario, the imagined Einstein could reject a reply that tests judge faithful to his earlier texts. That rejection would not prove the texts were misread. It could show that a person changes, that they address an audience, or that the system confused one period with another. It could also show nothing at all, since the rejection itself would be invented.

The only usable version of disagreement is documentary. Two dated pieces may support incompatible formulations. The model must then

present them together, or signal that it chooses a period. If it fabricates a seamless synthesis, it does not resolve the tension. It conceals it.

This is also where vocabulary becomes a risk. Describing the system as if it recognized, regretted or thought can make us forget the mechanisms under evaluation [SRC-025]. The book may use those verbs inside a clearly announced thought experiment. It cannot let them slide into a historical account.

What this part establishes

At the end of these four chapters, a conditional answer is possible. An assistant could organize a corpus, retrieve passages, distinguish technical regimes and submit certain regularities to tests. It could fail in instructive ways. It could neither make historical Einstein speak, nor turn imaginary recognition into a certificate.

If one cannot ask historical Einstein to recognize his model, **which independent criteria allow its fidelity to be measured?** Chapter 6 prepared the tools: accuracy, provenance, temporality, robustness, abstention, contamination. The present chapter explains why they are indispensable. Without them, the book's title would close on an impression.

This limit opens the next part. To understand what such an assistant would actually do, it is no longer enough to describe one particular corpus. One must explain how a model learns from data, what its parameters retain, and why a reply is not a record extracted from a database.

KEY POINTS

- 01 Recognition by historical Einstein cannot be collected.
- 02 Stylistic, documentary, historical, scientific and behavioral fidelity must not be confused.
- 03 A resemblance may come from a stereotype more than from a dossier.
- 04 The criteria of chapter 6 replace the impossible judge.
- 05 What follows explains the technical mechanisms that make those criteria necessary.

Sources used

- **SRC-025 - Shanahan.** Abstract and conclusion excerpts; caution on mental vocabulary applied to models.

- **SRC-030 and SRC-031.** Collective simulation results and distortions; limited analogy, not a direct transfer to Einstein.
- **SRC-040 - Norton.** Introduction consulted; the four named myths serve to identify discovery stereotypes, not to summarize every case study in the article.

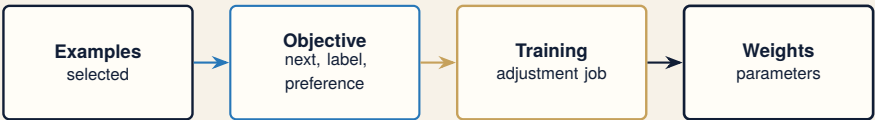
8 · From data to representations

A folder of examples is not the model. Training turns examples into weights, according to an objective.

The folder is not the skill

You can set a thousand pages on a table. They remain a thousand pages. Someone who has read them can then answer a question without reopening every sheet. The folder and the acquired skill are not the same object.

A machine-learning system likewise separates the examples from what is retained from them. The examples serve during a preparation phase. The usable result is a set of numerical values, the parameters, or weights. Once that phase is over, answering a new question is not reopening the folder page by page.



Explanatory sketch: the example set is not the resulting model.

Figure. Selected examples and a learning objective feed a training run. The usable result is a set of weights, not the folder itself. Author’s explanatory diagram. [SRC-010] [SRC-039], registry review levels.

This image has a limit. A person can justify an answer by finding the passage again. A model can produce a sequence of words without being able to show the page it would come from. The pedagogical resemblance stops there.

What the diagram lets us check

Three questions follow at once. Which examples were retained? What objective guided the adjustment? Were the resulting weights evaluated on

items held aside? Without these three answers, “the model has learned” remains a vague formula.

Adjusting is not conversing

The supervised fine-tuning documentation consulted on 9 October 2026 describes a precise operation. One assembles examples. One uploads them. A training job then updates the weights [SRC-039]. The same page indicates that the availability of this platform is shrinking. The description must therefore remain dated.

The abstract of the article by Ouyang and colleagues presents another sequence, for InstructGPT models: fine-tuning on demonstrations, then learning from human preferences [SRC-010]. That abstract concerns a family of models and a distribution of instructions. It also notes that simple errors can persist. It does not say that an ordinary conversation later restarts that training.

■ ESTABLISHED

In the cited documents, updating the weights proceeds through examples and a training job, or through the procedure described for InstructGPT. It is not presented as the immediate effect of an isolated question.

One architecture is not every system

The abstract by Vaswani and colleagues, in 2017, describes the Transformer. The proposed architecture relies on attention mechanisms. In that proposal, it does without recurrence and convolution [SRC-002]. The reported results concern translation first. The abstract alone does not describe the conversational assistants available in 2026.

It is therefore useful, and insufficient. Useful, because it shows that a computed representation can arise from a precise architecture, and not from a library of phrases stored as they are. Insufficient, because later systems add data, objectives, adaptations, and tools that this article does not catalog.

■ UNKNOWN

The 2017 abstract does not authorize describing the internal workings of every current commercial model. That description would still have to be established product by product.

What error allows us to say

Training can be explained without an equation. The pedagogical chain is:

selected examples → learning objective → error computation → parameter adjustment → evaluation on distinct examples.

The system produces a provisional answer. It is compared with an objective fixed in advance: the expected continuation, the label, an evaluator's preference. The gap is used to modify the weights. This presentation is a working image. It does not replace reading a methods paper, and none of the abstracts drawn on here has been treated as a line-by-line demonstration.

Pedagogical example: classifying a documentary genre

Take up again the task from chapter 6, without claiming to have run it. One assembles examples of notes, letters, and public texts. The objective is to estimate a genre. After fine-tuning, one evaluates on items held aside. If the examples contain almost only letters, the system may succeed well on letters and fail on notes. The result depends on the examples, the objective, and the evaluation - not on volume alone.

The image still suffices to avoid a confusion. A network is not "a database that would have understood everything." It is the result of an objective, selected examples, and adjustments. Changing the objective, or changing the examples, can change what the weights make probable. Raw volume does not guarantee quality. None of the abstracts drawn on here says that more examples alone produce a better model.

The next chapter asks what one calls a foundation model when the same base is reused for very different tasks.

KEY POINTS

- 01 Examples, objective, error, parameters, and evaluation form a chain.
 - 02 Training examples and the resulting weights are not the same object.
 - 03 A documentary classification task shows why the result depends on the chosen corpus.
 - 04 The 2017 Transformer is not a portrait of every 2026 assistant.
-

Sources used

- **SRC-002 - Vaswani et al., 2017.** Notice and abstract. Attention architecture. Not a description of later conversational models.
- **SRC-010 - Ouyang et al., 2022.** Notice and abstract. Fine-tuning then preferences for InstructGPT. Scope limited to the study.
- **SRC-039 - OpenAI supervised fine-tuning.** Page consulted on 9 October 2026. Platform availability evolving.

9 · Foundation models

A base trained broadly, then adapted, is neither general intelligence nor a simple file of texts.

A name for a practice

In 2021, Bommasani and colleagues propose calling foundation models those models trained on broad data, then adapted to several tasks. Their report discusses possibilities and risks, among them homogenization: the same defects can reappear in many applications built on a common base [SRC-003]. The document is a dated synthesis. The excerpts consulted are not the entire report, and 2021 is not 2026.

The name is useful if it remains descriptive. It designates a way of building and reusing a system. It does not prove that a model “understands everything,” nor that two products sharing that name have the same corpus, the same objective, or the same limits.

■ ESTABLISHED

The 2021 report defines a category and flags risks, including homogenization. It does not measure the products commercialized five years later.

Two ways to adapt

An adaptation can modify the weights. That is the sense of the fine-tuning described for InstructGPT: demonstrations, then human preferences, on a family of models [SRC-010]. It is also the sense of the training job in the documentation consulted on 9 October 2026 [SRC-039].

Another adaptation leaves the weights in place and changes only what is supplied with the question. The abstract by Brown and colleagues indicates that the few-shot tasks evaluated for GPT-3 are carried out without gradient updates during those tasks [SRC-009]. Examples placed in the prompt orient the present answer. They do not, by that alone, constitute a

new training of the general model.

These two paths can coexist in a product. Confusing them makes debate impossible. “The model learned from me this morning” can mean: the prompt changed the answer. Or: a training job modified weights. Or again: a product memory will later reinject a preference. Chapter 11 will separate these cases. Here it is enough to see that adaptation is not a single gesture.

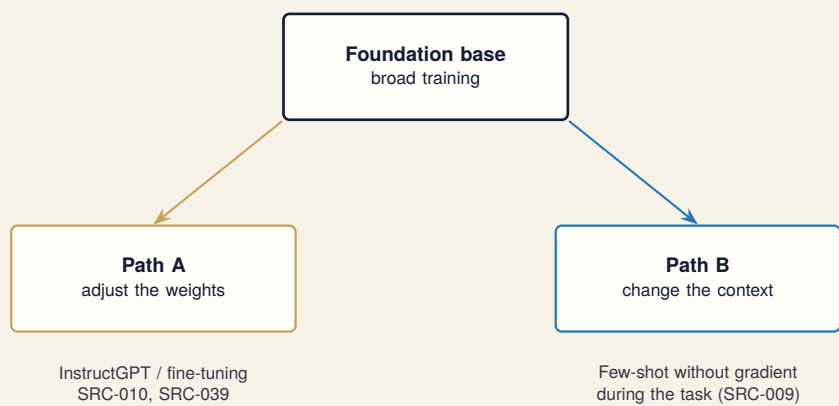


Figure. Two distinct adaptations of a foundation: change the weights, or only the context of the reply. Explanatory diagram. [SRC-003] [SRC-009] [SRC-010] [SRC-039].

Reading the diagram

Path A leaves a trace in the parameters. Path B leaves a trace in the message. The same screen can take one, the other, or both. As long as the product does not say which, the phrase “the model has learned” designates nothing inspectable.

Layer	What it provides	What it does not guarantee
Foundation base	Reusable general capabilities	Uniform quality on every task
Question context	Orientation of the present reply	Weight updates
Parameter tuning	Adaptation on examples	Coverage of all uses
Tools and external sources	Access to documents or services	Truth of the retrieved document
Own evaluation	Local measure of one use	Automatic transfer to another use

Table. Reusing a foundation is not deploying the same system. [SRC-003] [SRC-009] [SRC-010].

Reusing a base is not equivalent to deploying the same system. Two tools founded on a similar family of models can differ in their data, adaptations, external tools, and quality criteria. Performance does not transfer automatically from one use to another.

What scale does not settle

GPT-3, in the article by Brown and colleagues, is an example of a model evaluated at large scale on tasks little prepared one by one. The abstract also discusses difficulties and a risk of contamination of the test sets [SRC-009]. Extending that result to all models, all languages, and all questions would be another claim, which the article does not carry.

The risk flagged by Bommasani and colleagues remains readable without a new figure. If many tools start from the same base, an error in that base can repeat wherever it is adapted [SRC-003]. That is not a fatality measured for 2026. It is a reason to ask, for each use, which base, which adaptation, and which evaluation were chosen.

For the book's thought experiment, the consequence is direct. An assistant that would speak "like Einstein" might owe its tone to a general base, to an adaptation on a corpus, or only to the documents placed in the question. As long as these parts are not named, resemblance does not say where it comes from.

KEY POINTS

- 01 "Foundation model" is a category proposed in 2021, with discussed risks.
 - 02 To adapt can mean changing the weights or only the context.
 - 03 Base, context, fine-tuning, tools, and evaluation form distinct layers.
 - 04 The scale of a result does not turn it into a law of all systems.
-

Sources used

- **SRC-003 - Bommasani et al., 2021.** Abstract and excerpts. Definition and risks, including homogenization. Report not read in full.
- **SRC-009 - Brown et al., 2020.** Notice and abstract. Few-shot without gradient updates during the described tasks.
- **SRC-010 - Ouyang et al., 2022.** Notice and abstract. Scope limited to InstructGPT and the study's instructions.
- **SRC-039 - Supervised fine-tuning.** Page consulted on 9 October 2026.

10 · ChatGPT is not a database

A fluent answer can come from adjusted weights, a retrieved document, or both. It is not the same operation.

What this chapter adds

The previous chapter situated foundation models. This one isolates a single confusion: taking a fluent sentence for a retrieved record. The next chapter will fix a stable vocabulary. The one after that will turn to consent and the destinations of a conversation.

Retrieving is not continuing

A database, in the ordinary use of the word, keeps records and can return one when queried. One can mistype a key. One can receive an incomplete record. One does not, in principle, receive a sentence invented to fill the gap.

A language model does something else when it continues a text from its parameters. The produced continuation may resemble an exact memory. It may also be only probable given the training. Fluency does not say which of the two cases one has before one's eyes.

The abstract by Lewis and colleagues draws a distinction that still serves. A parametric memory is what the weights encode. A non-parametric memory is a set of documents that the system can retrieve and supply to the model when answering [SRC-004]. Their article describes a precise experimental architecture. It does not describe every product that has since claimed to "do search."

■ ESTABLISHED

The Lewis et al. abstract separates knowledge carried by the parameters from documents retrieved on demand. That separation alone does not describe every commercial product.

Three sentences that do not say the same thing

"The model knows it." One then asserts that the weights make this answer probable, without a document added for the occasion.

"The system retrieved it." One asserts that an external passage was searched for, then placed in the context of the answer.

"It was shown in the question." One asserts that the example is part of the current message. In the few-shot tasks described for GPT-3, that gesture changes the answer without a gradient update during evaluation [SRC-009].

A single screen can carry out one of these three operations, or combine them, without the final sentence bearing the label. Asking for provenance is not an expert refinement. It is the only way to know what was used.

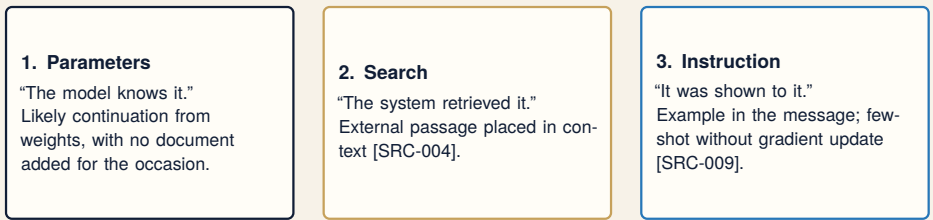


Figure. Three possible provenances of a fluent reply. Without labels, a single screen conflates them. Conceptual diagram. [SRC-004] [SRC-009].

Situation	What the system does	Needed check
Document search	Retrieves a passage	Document, page, context
Parametric generation	Produces a likely continuation	Every important claim
Generation with search	Combines documents and generation	Fidelity of the reply to sources

Table. Three distinct operations behind one window. Search is not a guarantee of truth. [SRC-004] [SRC-009].

The same question, three operations

Take up again the question from chapter 4 on how a formulation evolves between two versions. Three possible treatments:

1. **Retrieve** the two passages in the corpus and compare them. 2. **Generate** an answer from the regularities of the parameters alone. 3. **Generate after search:** place retrieved passages in the context, then produce a synthesis.

Verification is not the same. In the first case, one checks the document. In the second, each claim. In the third, the fidelity of the synthesis to the

sources. Document search is never a guarantee of truth: an incorrect, out-of-context, or poorly read source can also lead to error.

Why the distinction matters for Einstein

If the assistant "cites" a letter from 1905, one must know whether a volume was retrieved, whether an excerpt was pasted into the question, or whether a probable continuation was invented from the weights. The three cases do not deserve the same historical credit.

What a convincing answer does not prove

An answer can be right because the retrieved document was. It can be wrong because the document was, or because no source was consulted and the probable continuation filled a hole. It can be right by coincidence. This book's register does not contain, at this stage, a measure of ChatGPT's error rate in 2026. It would be dishonest to invent one.

The consequence for Einstein is clearer than any percentage. If the assistant cites a letter, one must be able to find the letter in an identified edition. If the assistant only "sounds right," one has a surface imitation. Earlier chapters already refused to take that imitation for an identity. This chapter adds the mechanism: without distinguishing parameters from retrieved documents, one does not even know what kind of imitation one is judging.

KEY POINTS

- 01 Search, generation, and search-assisted generation are three operations.
 - 02 Continuing a text from the weights is not returning a stored record.
 - 03 Search is not a guarantee of truth.
 - 04 No fluent sentence alone carries its provenance.
-

Sources used

- **SRC-004** - Lewis et al., 2020. Notice and abstract. Parametric / non-parametric distinction. Not all later products.
- **SRC-009** - Brown et al., 2020. Notice and abstract. No gradient update during the described few-shot tasks.

11 · Context, memory, and parameters

Four words must stay stable for the rest of the book: parameters, context, product memory, search.

The vocabulary that was missing

Chapters 8 to 10 showed mechanisms. This one does not re-demonstrate them: it names them once and for all. The parts that follow will have to use these four slots, or start the confusion again.

Four terms suffice, provided they are not swapped.

The **parameters** are the weights that result from prior training. Modifying them requires work on examples, as described in the supervised fine-tuning documentation consulted on 9 October 2026 [\[SRC-039\]](#).

The **context** is what is supplied for the current answer: the question, the instructions, the examples pasted into the message, the passages a tool has just added.

A **product memory** keeps certain information to reinject later, according to the service settings. The ChatGPT documentation consulted the same day separates these memory controls from the controls that concern the use of conversations to improve models [\[SRC-006\]](#) [\[SRC-007\]](#).

An **external search** retrieves documents and places them in the context. That gesture alone does not rewrite the weights [\[SRC-004\]](#).

■ ESTABLISHED

In the pages consulted on 9 October 2026, remembering a preference, retrieving a document, and adjusting weights are three distinct operations. That separation describes a product and a date, not a law of all systems.

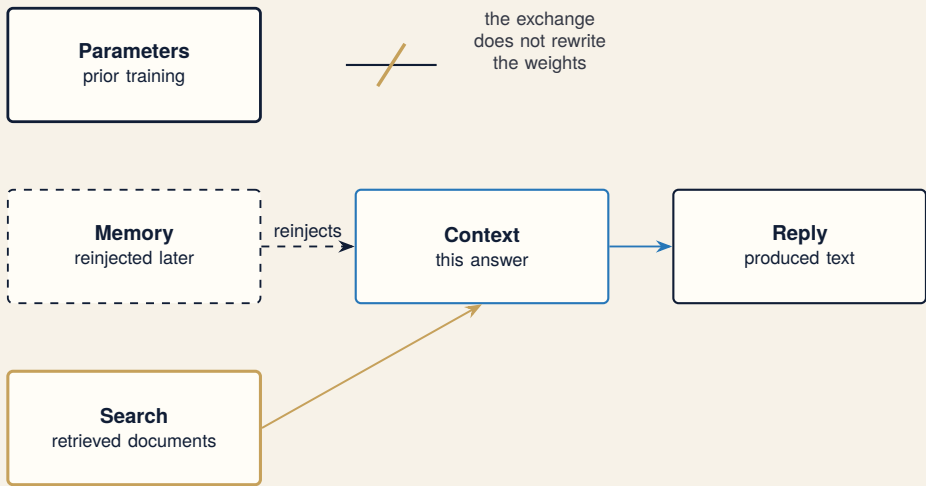


Figure. Parameters, context, product memory, and external search are not the same operation. The exchange feeds the context. It does not update the weights. Explanatory diagram, not an experimental result. [SRC-006] [SRC-007] [SRC-009] [SRC-039], at the registry’s review levels, product documentation dated 9 October 2026.

Mechanism	What changes	What does not change
Parameters	Weights after a training run	An isolated question in the window
Context	The current reply (message, tools)	Weights of the general model
Product memory	What may be reinjected later	A person’s autobiography
External search	Documents placed in the context	Parameters, by that act alone

Table. Four mechanisms not to be swapped. Descriptions dated for product documentation consulted on 9 October 2026. [SRC-004] [SRC-006] [SRC-007] [SRC-039].

What a conversation can and cannot do

A sentence sent in a window enters a context. It can change the next answer. In the few-shot tasks evaluated by Brown and colleagues, that change occurs without a gradient update during those tasks [SRC-009].

The same sentence does not become, by being sent, a new training of the general model. The data-controls documentation distinguishes history from choices related to improving models [SRC-006]. A policy may provide for later use. That use would occur in another timeframe, with other controls. It is not to be confused with the answer one has just read.

Temporary chat, in the page consulted on 9 October 2026, is presented as a mode without product memory and without declared use for training. The page also notes security retentions and exceptions [SRC-008]. “Temporary” is therefore not a synonym for “nothing is kept, in any sense.”

■ ESTABLISHED

A conversation must not be described as an instantaneous modification of the general model's parameters. Retention exceptions and later policies must remain named, and dated.

Diagnosis: "the assistant seems to remember"

When an answer gives the impression of a memory, four explanations must be examined before any conclusion:

1. **Still-accessible context** - the information appears in the message or in the history of the current conversation.
2. **Product memory** - a preference or a fact was stored and reinjected according to the service settings.
3. **External search** - a document was retrieved and placed in the context.
4. **Parametric regularity** - the weights make this continuation probable without a document added for the occasion.

The appearance of the answer alone does not decide the mechanism. This diagnostic sequence is the book's reference map. Later chapters should return to it rather than redefine the same words.

Why the book stops on these four words

The whole Einstein scenario depends on this grid. A historical corpus placed in a search is not an autobiographical memory. A style preference kept by the product is not a notebook from 1905. Adjusting weights on edited letters would not be a conversation with the author of those letters. And none of these gestures would make present a man who died in 1955.

The parts that follow may speak of human traces, world models, or loops between generated texts and new training. They will do so with these four slots, or they will restart the confusion this part has just undone.

Verification itself is not closed. Product pages change. Article abstracts are not the articles. The note placed before the prologue says it for the whole book: a convincing answer does not stand in for proof. Here it applies to the vocabulary itself. If a documentation evolves, the corresponding slot must be rewritten, not defended by habit.

KEY POINTS

- 01 Parameters, context, product memory, and external search form the reference map.
 - 02 When a system "seems to remember," four explanations must be examined.
-

03 A question first changes a context; it does not update the weights by itself.

04 This grid will be reused, not redefined, in the rest of the book.

Sources used

- **SRC-004** - **Lewis et al.** Notice and abstract. Retrieval distinct from parametric memory.
- **SRC-006** - **OpenAI data controls**. Page consulted on 9 October 2026.
- **SRC-007** - **OpenAI memory**. Page consulted on 9 October 2026. Features vary by offering.
- **SRC-008** - **Temporary chat**. Page consulted on 9 October 2026. Security retentions not to be forgotten.
- **SRC-009** - **Brown et al.** Notice and abstract. No gradient update during the described few-shot tasks.
- **SRC-039** - **Supervised fine-tuning**. Page consulted on 9 October 2026.

12 · Is every conversation data?

A sentence once sent exists. It becomes training data only if a processing chain decides so.

The question

Chapter 11 fixed four mechanisms. Here the question changes: not “how the answer is computed,” but **where the trace goes once it is sent**. This is a problem of governance and consent.

Flagged fictional case

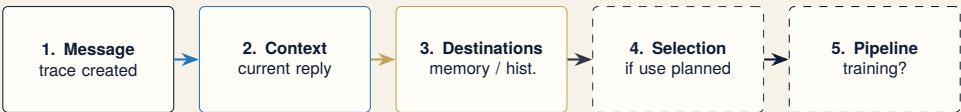
■ **THOUGHT EXPERIMENT**

The following case is fictional. It serves to make possible destinations visible, not to describe a real use by a named user.

A user asks for a synthesis. He corrects the answer. Later, he finds a format preference again. Three distinct events may have occurred: the message influenced the context; a preference was remembered; nothing yet indicates that a training pipeline used that exchange. Confusing the three is believing that “every conversation immediately trains the model.”

Conditional lifecycle

Between typing and a possible reuse, several steps may intervene: processing the answer, retention, access control, filtering, selection, preparation, then - only if an appropriate pipeline exists - training or evaluation.



Conditional branches by product, settings, and policies. Not a law for all services.

Figure. Conditional lifecycle of a conversation: message, context, destinations, selection, possible pipeline. Conceptual diagram; depends on product and policies. [SRC-006] [SRC-007] [SRC-008].

Function	Question	What it does not prove
Active context	Shapes the current reply?	Lasting weight learning
Product memory	Later reuse possible?	Change of general parameters
Document search	Content retrieved on demand?	Truth of the content
Training	Examples used in a pipeline?	Integration of every conversation

Table. One trace, four functions. Distinguish technically possible, actually implemented, and allowed by policy.

The ChatGPT documentation consulted on 9 October 2026 distinguishes memory controls from those that concern the use of conversations to improve models [SRC-006] [SRC-007]. Temporary mode is presented without product memory and without declared use for training, with security retentions [SRC-008]. These pages describe a product and a date, not a law of all systems.

■ ESTABLISHED

In the pages consulted on 9 October 2026, product memory and use for improving models are separate controls.

Possible, implemented, authorized

Three levels must still be distinguished. **Technically possible:** an architecture could retain or reuse a trace. **Actually implemented:** a service does so, for certain data, at certain dates. **Authorized by a policy:** settings, consent, retention duration, and deletion allow it. Without these three levels, the word "data" erases the decision.

Consent, confidentiality, duration, deletion, and traceability are not annexes. They decide whether a conversation remains a context, becomes a product memory, or enters - possibly - a pipeline.

Consequence

Einstein's imaginary exchanges with an assistant would not automatically be a learning corpus. Without a processing record, the scene says nothing about what the system would learn. The reader must be able to refute the simplistic claim of immediate training at every message.

KEY POINTS

- 01 A conversation leaves traces; it does not train automatically.

-
- 02 Context, memory, search, and training are four functions.
-
- 03 Possible, implemented, and authorized are not to be confused.
-
- 04 Without a named destination, the word "data" remains too broad.
-

Sources used

- **SRC-006 - Data controls.** Page consulted on 9 October 2026.
- **SRC-007 - Memory.** Page consulted on 9 October 2026.
- **SRC-008 - Temporary chat.** Page consulted on 9 October 2026.
- **SRC-009 - Brown et al.** Notice and abstract. Few-shot without gradient updates during the described tasks.

13 · Does the researcher feed the AI?

Publications, demonstrations, and human preferences can enter a learning chain. They do not all enter in the same way.

The question

Does the researcher feed the AI? The wording suggests a destiny. The real question is: **through which mediations can a scientific production modify the materials, criteria, or behaviors of a system?**

A case in three stages

■ THOUGHT EXPERIMENT

Illustrated conditional chain. No automatic entry into training is asserted.

1. A study is published with method, results, and limits. 2. An AI produces a clear synthesis of it. 3. That synthesis is taken up in a course or a secondary article.

Each mediation can preserve, simplify, distort, or lose information. A generated output does not automatically enter a future training. It can, however, become a new public trace - and, if a pipeline collects it, a candidate datum.

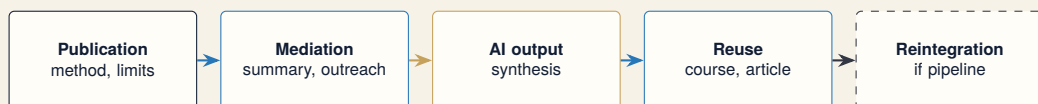


Figure. From publication to possible reintegration. Each mediation may preserve, simplify, or distort. Author's conditional diagram.

Element	Risk	Control
Result	Overgeneralization	Return to the original article
Method	Omitted step	Protocol check
Limit	Lost in the summary	Keep the caveats
Error	Amplification	Provenance and validation
Uncertainty	Turned into certainty	Confidence levels

Table. What passes from a publication to a synthesis. A generated output does not automatically enter training.

Articles, abstracts, comments, outputs

Statuses must be separated. The original article carries the method and the reservations. The abstract may set them aside. The comment adds an interpretation. The generated output may turn an uncertainty into a certainty. Messeri and Crockett analyze the risk of illusions of understanding in science [SRC-013]. A tool that speeds up a task can make one believe it has grasped the field.

The abstract by Ouyang and colleagues describes, for InstructGPT, demonstrations and human preferences as fine-tuning signals [SRC-010]. That is neither a human reading an article, nor an ordinary conversation.

■ ESTABLISHED

The InstructGPT abstract describes demonstrations and human preferences as sources of fine-tuning. It does not describe every scientific use of AI.

Why clarity can mislead

A researcher may prefer an answer because it is clear. Clarity is not correctness. If the error is then taken up without returning to the source, the chain amplifies an appearance of knowledge. The minimal check remains returning to the original article and keeping the limits.

Consequence

Feeding a model scientific texts is not transmitting the whole craft. It is giving it selected traces - and sometimes already distorted mediations. The next chapter asks what these traces let one see of methods.

KEY POINTS

- 01 Publishing makes a trace available; that alone does not decide its use.

-
- 02 Each mediation can lose a limit or amplify an error.
 - 03 Clarity and correctness are not to be confused.
 - 04 A generated output does not automatically enter a training.
-

Sources used

- **SRC-001 - Einstein Papers Project.** Scope of an edited corpus, not mental exhaustiveness.
- **SRC-010 - Ouyang et al.** Notice and abstract. Demonstrations and preferences.
- **SRC-013 - Messeri and Crockett.** Notice and abstract. Illusions of understanding in science.

14 · Does AI also learn our methods?

Texts can show procedures. They do not always show why they were chosen, or what was set aside.

Method in the text

A demonstration, a written proof, a protocol, or commented code makes a way of doing visible. A model trained on such texts can reproduce certain regularities: the order of steps, a discipline's vocabulary, the form of a demonstration. The InstructGPT abstract shows that a system can be fine-tuned to follow instructions formulated by humans better [\[SRC-010\]](#). Following an instruction is not yet mastering a scientific method.

Norton recalls, in the introduction read for this book, that popular discovery narratives sometimes simplify long inquiries [\[SRC-040\]](#). If a corpus contains mostly these narratives, a model may learn the simplification rather than the method. Stylistic resemblance is not fidelity to the work.

Documentary case study - three questions

Take up again the grid from chapter 1 on a historical dossier verified in categories, not in invented intentions. The notices of the *Collected Papers* already separate genres and periods [\[SRC-033\]](#) [\[SRC-035\]](#) [\[SRC-036\]](#).

1. **Observables.** Which changes of formulation are directly visible between dated versions? 2. **Cautious inferences.** Which orders of operations or types of corrections can one compare without claiming to read intention? 3. **Unknowns.** Which motivations, unwritten hesitations, or laboratory contexts remain outside the records?

Do not attribute to Einstein a psychological intention that the records do not allow one to establish. Learning produced knowledge and learning regularities of method are two distinct objects.

What demonstrations omit

A successful demonstration shows a path. It shows less the unpublished errors, unavailable instruments, team negotiations, or local norms. Premack and Woodruff already defined the attribution of mental states as an inference about what is not directly observable [SRC-027]. Applied to methods, the same caution holds: one infers a procedure from traces. One does not observe the whole intention.

■ UNKNOWN

None of the abstracts drawn on here measures the share of scientific method actually captured by a contemporary language model.

Consequence for Einstein

If one wanted an assistant to “learn Einstein’s method,” one would first have to specify which: drafting an article, correcting a note, an epistolary exchange, or a later narrative of a discovery. The notices of the Collected Papers already separate document genres [SRC-033] [SRC-035] [SRC-036]. Confusing them would produce an average method that would have existed nowhere.

The gain of this part is therefore simple. A human trace can become data. It becomes so through choices. And what it teaches depends on what it shows, what it hides, and the learning objective.

KEY POINTS

- 01 Texts can encode procedural regularities without transmitting the whole craft.
 - 02 Simplified narratives can be easier to imitate than long inquiries.
 - 03 Attributing a method remains an inference from traces.
 - 04 Documentary genres are not interchangeable.
-

Sources used

- **SRC-010 - Ouyang et al.** Notice and abstract.
- **SRC-027 - Premack and Woodruff.** Notice and abstract. Inference about unobserved states.
- **SRC-033, SRC-035, SRC-036 - Volume notices.** Documentary genres, not individual pieces.

- **SRC-040 - Norton.** Introduction and framing. Discovery myths.

15 · What is a “world model”?

The same word designates different objects in cognitive science, in reinforcement learning, and in everyday AI talk.

An overloaded word

“World model” suggests a small inner copy of reality. The image is strong. It is also dangerous. A representation can serve to predict or to act without resembling the world as a map resembles a landscape.

The editorial presentation of Johnson-Laird describes mental models as structures that make it possible to reason about relations among objects and events [SRC-018]. This wording comes from a publisher’s notice. The full work was not read for this book. It suffices to recall that in cognitive science, “model” is not synonymous with an inner photograph.

A precise use in learning

Ha and Schmidhuber call “world models” generative models of environments, used especially to learn in reinforcement-learning settings [SRC-011]. The abstract describes compressed spatiotemporal representations. It does not allow one to conclude that every language model possesses, in the same sense, a world model.

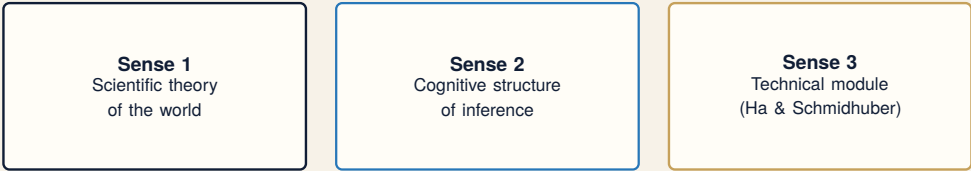
■ ESTABLISHED

In the Ha and Schmidhuber abstract, “world model” designates a precise learning framework. That framework does not extend automatically to all LLMs.

Three senses not to merge

The first sense is descriptive: a scientific theory of the physical or social world. The second is cognitive: an internal structure that guides inference or action. The third is technical: a module that predicts states of an environment for an agent. A conversational assistant can excel at produc-

ing sentences about the world without one’s having shown which of these three senses applies.



Three distinct objects. Merging them turns a metaphor into proof.

Figure. Three senses of “world model”. The book requires a qualifier before using the phrase. Conceptual diagram. [SRC-011] [SRC-018].

Sense	Object	Relevant question
Scientific theory	Explanation of a phenomenon	Which observations support it?
Cognitive model	Structure for reasoning	Which inferences does it allow?
Technical model	State prediction	Which predictions are correct?
LLM representation	Debated learned structures	Which tests characterize them?

Table. Four senses of the word “model”. Not equivalent. [SRC-011] [SRC-018].

Criterion of use in the book

For what follows, we will use “world model” only with a complement: world model in the sense of Ha and Schmidhuber, representation useful for a task, or popular discourse. Without that complement, the next chapter could not correctly pose the question of what a system actually possesses.

A simple test helps the reader. If one can replace the phrase with “prediction of states for an agent,” one is in the technical sense. If one means “scientific theory,” one must say so. If one means “human mental structure,” one needs a cognitive source, not a chatbot demonstration.

KEY POINTS

- 01 “World model” has several disciplinary senses.
- 02 Johnson-Laird is cited here via a publisher’s notice, not via a full reading.
- 03 Ha and Schmidhuber define a limited technical use.
- 04 Merging the senses turns a metaphor into proof.

Sources used

- **SRC-011** - **Ha and Schmidhuber**. Notice and abstract.
- **SRC-018** - **Johnson-Laird**. Editorial presentation only.

16 · Does AI possess a representation of the world?

Performances and debates exist. A single verdict on understanding is not established by this book's sources.

What one can measure

One can measure what a system predicts, classifies, or generates. One can compare these outputs with references. One can sometimes inspect internal states. None of these operations, by itself, decides whether the system “possesses a representation of the world” in the sense in which a human speaks of it.

Mitchell and Krakauer map the debate on the understanding of large language models and insist on the need for differentiated criteria [\[SRC-019\]](#). Their perspective article does not supply a single experimental verdict. Bender and Koller argue, in a position paper, that training on linguistic form alone does not, by itself, establish the learning of meaning [\[SRC-020\]](#). That is an authors’ argument, not a measure of all later multimodal systems.

■ UNKNOWN

The sources consulted do not authorize concluding that contemporary language models do, or do not, possess a representation of the world in the human sense.

Five levels of evaluation

Before speaking of a world representation, one must know which test was run.

Level	Question	Allowed conclusion
Performance	Task success?	Success under these conditions
Generalization	New cases?	Generalization within scope
Robustness	Reformulations?	Relative stability
Relations	Causal or temporal links?	Some relations hold or not
Mechanisms	Testable internal cues?	Hypotheses about representations

Table. Five evaluation levels. None alone measures human understanding. [SRC-019] [SRC-020].

None of these levels constitutes, by itself, a measure of human understanding or of consciousness. The reader must refuse a conclusion stronger than the test justifies.

What one must not confuse

Ha and Schmidhuber describe world models in a reinforcement-learning framework [SRC-011]. That framework is not ChatGPT. Shanahan recalls, in a synthesis article, the need for a language that reflects the mechanisms of models and avoids unjustified imputations [SRC-025]. Attributing global understanding because an answer seems informed is precisely the imputation to avoid.

A representation useful for a task can exist without subjective experience, without intention, and without a complete map of the world. The book has already said so for a person. It repeats it for a system.

Consequence for the inquiry

If Einstein’s assistant answers a physics question correctly, one has a performance. If the answer cites a retrieved document, one has a provenance. If the weights make the continuation probable, one has an encoded regularity. None of these cases alone settles the debate on understanding.

KEY POINTS

- 01 Measuring a performance is not equivalent to settling understanding.
- 02 Mitchell and Krakauer map a debate, without a single verdict.
- 03 Bender and Koller formulate a position argument on form and meaning.
- 04 Vocabulary must stick to the mechanism, not to the impression.

Sources used

- **SRC-011** - Ha and Schmidhuber. Notice and abstract.
- **SRC-019** - Mitchell and Krakauer. Notice and abstract.
- **SRC-020** - Bender and Koller. Abstract and conclusion excerpts.
- **SRC-025** - Shanahan. Abstract and conclusion excerpts.

17 · From millions of traces to a collective representation

Aggregation makes certain regularities visible. It also masks absences, languages, and selection choices.

The question

What does “collective” mean here? Not a shared consciousness. An **aggregation of human productions selected, transformed, and weighted** by technical and institutional mechanisms.

Case: repetition is not proof

A frequently repeated claim can dominate a corpus. A hundred documents that repeat the same study do not constitute a hundred independent confirmations. Argyle and colleagues show that a conditioned model can reproduce certain response distributions [\[SRC-030\]](#). Aher and colleagues also observe a distortion of over-precision [\[SRC-031\]](#). Aggregation can smooth as much as it reveals.

■ ESTABLISHED

In the cited studies, aggregation can reproduce certain collective distributions. It does not demonstrate the restitution of a person.

Map of absences

Before celebrating volume, list possible absences: languages, periods, regions, documentary genres, scientific fields, social groups. Bommasani and colleagues underline risks of homogenization for foundation models [\[SRC-003\]](#). An oral tradition, unpublished knowledge, or an under-documented population does not magically enter the representation.

Einstein’s archives recall the same principle at another scale. An editorial project makes categories accessible. It does not make accessible every-

thing that took place [SRC-001] [SRC-032].

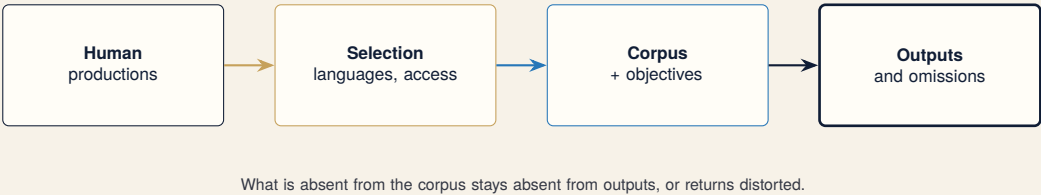


Figure. Building a collective representation: productions, selection, corpus, outputs, and omissions. Conceptual diagram. [SRC-003].

Situation	What one might believe	To check
Much English	World knowledge covered	Other languages
Many retellings	Independent consensus	Primary sources
Few documents	Marginal phenomenon	Conservation obstacles
Many digitized items	Complete archive	Gaps and non-digitized material

Table. Abundance versus representativeness. A hundred retellings of a study are not a hundred confirmations.

Diversity and independence

Diversity of sources and independence of sources are not to be confused. Translation, indexing, commercial availability, and access rights also decide what becomes visible. A useful collective representation must be presented as such: an estimate on a corpus, for a task, with nameable biases.

Consequence

The collective representation is not a neutral photograph of humanity. It depends on what has been preserved, selected, made accessible, and evaluated. The next chapter organizes conditional futures - not a single trajectory.

KEY POINTS

- 01 "Collective" designates a selected aggregation, not a consciousness.
- 02 Abundance and representativeness diverge.
- 03 A hundred reprises do not equal a hundred independent proofs.
- 04 A corpus's absences reproduce themselves in the outputs.

Sources used

- [SRC-001](#), SRC-032 - Einstein Papers Project. Scope and institutional history.
- [SRC-003](#) - Bommasani et al. Abstract and excerpts.
- [SRC-030](#) - Argyle et al. Notice and abstract.
- [SRC-031](#) - Aher et al. Notice and abstract.

18 · 2026 → 2030 → 2040 → 2050

Conditional scenarios help thinking. They are not forecasts.

■ **THOUGHT EXPERIMENT**

The dates in this chapter predict no technological calendar. They organize possible conditions.

Why horizons, and not oracles

A thought experiment explores consequences. It does not turn a scenario into an event [SRC-038]. Saying “in 2040, systems will do X” without conditions is dressing a preference as a chronology.

Three conditional scenarios

Scenario A - Managed integration. Better-documented corpora, independent evaluations, trusted institutions. Condition: durable funding, governance, access to sources.

Scenario B - Fragmented progress. Rising capabilities, but inequalities across languages, domains, and access rights. Condition: unevenly distributed investment.

Scenario C - Degraded loops. Generated content reused without sufficient provenance; repeated errors; verification lagging. Condition: production speed and insufficient validation.

A · Mastered Traceability, audits, verifiable sources. Condition: governance.	B · Fragmented Uneven capabilities by language and access. Condition: investment.	C · Degraded Reuse without provenance. Repeated errors. Condition: speed.
---	---	--

Figure. Three conditional scenarios, not an inevitable chronology. Discussion diagram. [SRC-038].

Variable	Indicator	Why
Documentary quality	Sourced claims	Traceability
Diversity	Languages, regions, domains	Visible absences
Reliability	Errors on independent tests	Comparison
Governance	Audit and redress	Correction possible
Dependence	Infrastructure concentration	Power

Table. Proposed indicators for tracking conditional scenarios. Research proposals, not measures obtained.

These indicators are research proposals, not measures already obtained. If the share of sourced claims falls while synthetic content dominates without a filter, scenario C gains plausibility. If independent audits and effective remedies become widespread, scenario A becomes more credible.

Horizon	Plausible mechanism	What would invalidate it
2026	Distinctions parameters / context / memory / search	Treating a product policy as a timeless law
2030	Denser tool integration	Claiming a conversation equals training
2040	Generated-content loops ↔ corpora	Ignoring homogenization risks already discussed
2050	Collective-memory infrastructures	Equating collective archive with autobiographical memory

Table. Conditional horizons, not a technology calendar. Methodological frame [SRC-038]; dated product documentation [SRC-006] [SRC-007].

■ **HYPOTHESIS**

The 2030-2050 horizons are discussion frames. None of the register’s sources fixes these dates as facts.

What the book refuses

It refuses the single curve of progress. It also refuses the total refusal to imagine. The next chapter asks what a generation of models can actually inherit from a previous one - through pipelines, not through a biology.

KEY POINTS

- 01 Three conditional scenarios replace a single chronology.
- 02 Each scenario has conditions and indicators.
- 03 The proposed indicators are not results obtained.

04 Without conditions, a date becomes an advertising narrative.

Sources used

- **SRC-003** - Bommasani et al. Risks discussed in 2021, not a calendar.
- **SRC-006**, SRC-007 - OpenAI documentation. Dated 9 October 2026.
- **SRC-038** - Thought experiments. Methodological frame.

19 · Does each generation inherit from the previous one?

Weights, corpora, evaluations, and human decisions are transmitted through pipelines, not through a natural genealogy.

Inheriting is not growing up

A generation of models can take up weights, distill a larger system, reuse a corpus, or train in part on prior outputs. Each of these gestures is an engineering choice. None equals a biological maturation.

Supervised fine-tuning, in the documentation consulted on 9 October 2026, starts from examples and a job that updates weights [SRC-039]. Instruct-GPT, in the abstract by Ouyang and colleagues, also starts from demonstrations and preferences [SRC-010]. In both cases, inheritance passes through data and objectives. It does not pass through a lived memory.

Matrix of inheritance mechanisms

Inheritance is not automatic. It results from engineering, selection, and governance decisions.

Mechanism	What is transmitted	Risk / control
Reused weights	Prior capabilities	Independent tests
Distillation	Source behaviors	Task comparison
Reused corpus	Documents and biases	Provenance
Synthetic data	Generated outputs	Independent references
Reused evaluations	Criteria and thresholds	Held-out sets

Table. Technical inheritance: engineering decisions, not biological maturation.

Useful skills can be transmitted. Errors too. If synthetic outputs enter a new corpus without a filter, a misleading regularity can strengthen. This book’s sources do not measure the scale of that effect in 2026. They only authorize naming it as a conditional risk [SRC-003]. Any generalization to “all models” or “all synthetic data” would go beyond what is established

here.

■ **HYPOTHESIS**

The transmission of errors via synthetic corpora is a plausible risk. Its intensity is not established here by a single measure.

Human decisions in the chain

Who chooses the corpus? Who filters? Who evaluates? Who decides that one system replaces another? The NIST AI risk-management framework, in its general scope consulted, insists on practices of reliability and governance, without by itself resolving every product [SRC-005]. Technical inheritance does not erase responsibility.

For Einstein, the question becomes: a model fine-tuned on a 2026 edition would inherit that edition, its choices and its silences, not the life of 1905. The next chapter examines the loop between humans and systems when outputs return into practices.

KEY POINTS

- 01 Inheritance between models passes through pipelines.
 - 02 Errors can be transmitted along with skills.
 - 03 Human decisions remain in the chain.
 - 04 Inheriting an edition is not inheriting a life.
-

Sources used

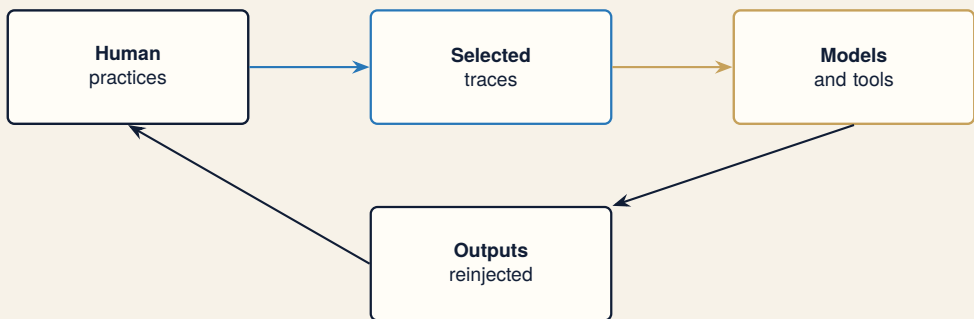
- **SRC-003** - Bommasani et al. Foundation risks.
- **SRC-005** - NIST AI RMF. General scope, voluntary framework.
- **SRC-010** - Ouyang et al. Notice and abstract.
- **SRC-039** - Supervised fine-tuning. Dated page.

20 · The AI ↔ humanity loop

Feedbacks exist as soon as outputs enter practices. They do not automatically improve what they touch.

A loop is not progress

Humans produce texts. Systems train on or rely on traces. Their outputs return into articles, code, conversations, and decisions. The loop is real as soon as that circulation exists. It is not, by itself, a guarantee of improvement.



The loop exists as soon as there is circulation. It is not automatic progress.

Figure. Practices, traces, models, and outputs form a loop once they circulate. This is not automatic progress. Conceptual diagram. [SRC-003] [SRC-005].

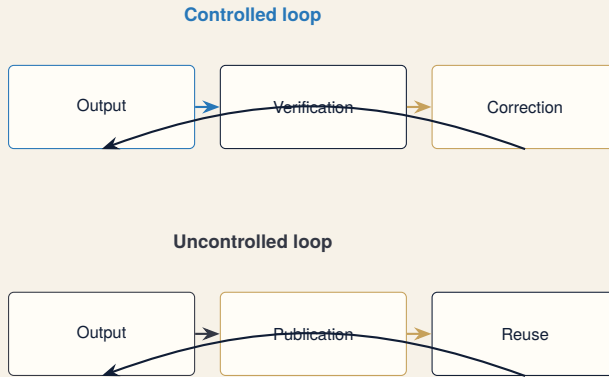


Figure. Controlled loop versus uncontrolled loop. An error may gain visibility without gaining truth. Conceptual diagram.

The report on foundation models already discusses opportunities and risks linked to widely reused systems [SRC-003]. A generalization that would say “more users, therefore better models” was set aside from the register as unsupported. Quality depends on selection, filtering, the objective, and evaluation.

Two loops, two fates

In a **controlled loop**, the output is verified, corrected, sourced, then possibly reused. In an **uncontrolled loop**, a plausible but false output is published, taken up as a source, then repeated. The error gains visibility without gaining truth. Social propagation and integration into a future training remain distinct: the first can occur without the second.

Control points in the loop

Filters can be placed at four points: selection of traces, training objective, evaluation of outputs, reintroduction into practices. A loop without these controls can amplify an appearance of knowledge as readily as verified knowledge.

■ ESTABLISHED

No source in the register supports that an increase in the number of users guarantees, by itself, improvement of models.

Filters and institutions

A loop can be filtered. One can refuse certain synthetic contents, document corpora, evaluate drifts, or leave humans to decide. The NIST framework recalls, in its general scope, the importance of risk-management prac-

tices [SRC-005]. Messeri and Crockett underline the risk of illusions of understanding in scientific uses [SRC-013]. Without filters, the loop can propagate an appearance of knowledge.

Return to Einstein

If texts generated “in the manner of Einstein” entered new corpora without a label, the next generation could learn the stereotype rather than the dossiers. Norton has already shown how discovery myths eclipse long inquiries [SRC-040]. The loop would then make the myth more fluent, not more true.

The gain of this part is institutional as much as technical. A loop exists through pipelines, choices, and organizations. It does not exist as a natural destiny of artificial intelligence.

KEY POINTS

- 01 The circulation of outputs creates possible feedbacks.
 - 02 A rise in use does not guarantee better quality.
 - 03 Filters and evaluations decide much of the loop.
 - 04 A stereotype can strengthen faster than an archive.
-

Sources used

- SRC-003 - Bommasani et al.
- SRC-005 - NIST AI RMF.
- SRC-013 - Messeri and Crockett.
- SRC-040 - Norton. Introduction.

21 · Can AI learn “who we are”?

Identity, preferences, habits, and roles do not form a single object. Predicting them is not possessing them.

Cutting up the question

“Who we are” gathers too many things. A narrative identity links a reconstructed past and an imagined future [\[SRC-026\]](#). Preferences can be stable or local. Habits leave digital traces. Social roles change with context. A system can estimate one without touching the others.

Kosinski, Stillwell, and Graepel show that certain attributes are predictable from defined digital behavioral traces, in a defined dataset [\[SRC-029\]](#). Youyou, Kosinski, and Stillwell observe that computer judgments of personality based on Facebook Likes outperformed some human judgments according to their study’s criteria [\[SRC-028\]](#). These results are bounded. Precision is not understanding of a person.

■ ESTABLISHED

In the cited studies, certain attributes are predictable from defined digital traces. That does not establish a restituted identity.

What a prediction authorizes one to say

A drafting preference observed in a corpus is not a complete personality. A strategy repeated in situations seen is not a law forever. A prediction conditional on a context is not knowledge of real convictions. Contradictory documents may signal an evolution or different sources - not “the true self” revealed by one of them. Counterexamples: change of mind, adaptation to an audience, gap between what is written and what is done.

Inferring is not reading

Human mentalizing infers mental states from cues [\[SRC-022\]](#). Premack

and Woodruff already formulated this attribution as a strategy for predicting behavior [SRC-027]. Dennett describes an intentional strategy useful for prediction, without making it a direct reading of the mind [SRC-024]. Systems that predict traits or simulate answers do, at best, work of the same kind: estimate for a task.

Tulving distinguished episodic memory and semantic memory [SRC-016]. Conway and Pleydell-Pearce describe autobiographical recollection as construction [SRC-017]. A product profile that retains a preference enters none of these human categories.

What the question authorizes

It authorizes saying that a system can learn useful regularities on traces. It forbids concluding that it knows who we are. The next chapter examines the limiting case of the intellectual twin.

KEY POINTS

- 01 Identity is not a single score.
 - 02 Bounded predictions exist in precise studies.
 - 03 Inferring a trait is not restituting a person.
 - 04 Human memory and product memory remain distinct.
-

Sources used

- SRC-016, SRC-017, SRC-022, SRC-024, SRC-026, SRC-027. Levels of examination in the register.
- SRC-028, SRC-029. Bounded empirical abstracts.

22 · The intellectual twin

A useful twin requires a task, consent, a test, and a right of revocation. It is not a double.

Define before imitating

An intellectual twin, in this book, designates a system meant to assist or simulate certain aspects of an identifiable intellectual production. The definition is operational. It asserts no ontological identity.

Park and colleagues build agents with natural-language memory, retrieval, reflection, and planning, in a sandbox of twenty-five agents [SRC-012]. Rabinowitz and colleagues train a network to predict artificial agents from observed behaviors [SRC-023]. Argyle and Aher explore simulations of samples or experimental results [SRC-030] [SRC-031]. None of these studies delivers a double of Einstein, or of a living consenting person.

■ **ESTABLISHED**

The cited simulations concern agents, distributions, or tasks. They do not establish a faithful personal twin.

Minimal conditions

Without a precise task, the twin becomes a slogan. Without period and corpus, it mixes eras. Without traces held aside for the test, it repeats what it was shown. Without consent, it turns a person into material. Without a right of revocation, it freezes a version the person can no longer correct.

Field	Requirement	Limit
Modeled object	Style, method, or prediction	Not the whole identity
Corpus	Periods, languages, exclusions	Visible selection
Evaluation	Criteria before tests	No abusive single score
Abstention	Signal insufficiency	Do not force a voice
Governance	Rights and responsibility	Do not speak “in the name of”

Table. Identity sheet for an intellectual twin. Operational definition.

Six dimensions without an abusive single score: documentary fidelity, temporal fidelity, quality of reasoning, robustness, abstention, transparency of sources. An excellent stylistic imitation does not compensate for invented citations.

Temporal drift is inevitable. A person changes. A historical corpus is frozen, or updated by others. A twin that does not display its date lies about what it represents.

For Einstein, and for us

The personal chatbot imagined earlier is only an extreme and posthumous case. For a living person, the stake is more immediate: who controls the model, who reads its outputs, who can shut it off. The next chapter asks what a library of scientific personalities would produce.

KEY POINTS

- 01 A twin is a situated construction, not a duplicate.
 - 02 Agent and simulation studies do not validate a personal identity.
 - 03 Task, corpus, test, consent, and revocation are minimal conditions.
 - 04 The model's date is part of its truth.
-

Sources used

- [SRC-012](#), [SRC-023](#), [SRC-030](#), [SRC-031](#). Experimental scopes in the register.

23 · A library of scientific personalities

A plurality of pedagogical models can illuminate. A simulated authority can mislead.

The possible pedagogical value

One can imagine assistants that present styles of reasoning, controversies, or corpora associated with scientific figures. Their interest would be to make readings, tensions, and documents accessible. Their danger would be speaking with the voice of an authority no longer there to correct itself.

The Collected Papers already organize a documentary plurality: writings, correspondence, notes [\[SRC-001\]](#) [\[SRC-033\]](#) [\[SRC-036\]](#). A library of personalities that flattened these genres into a single fluent voice would betray the edition it claims to extend.

The risk of simulated authority

Norton shows how discovery myths can eclipse long inquiries [\[SRC-040\]](#). A model trained on the most repeated versions of these myths would have the authority of tone without the discipline of the dossier. Shanahan recalls that language that is too mentalistic credits systems with capacities they do not establish [\[SRC-025\]](#).

■ HYPOTHESIS

A library of scientific personalities could have pedagogical value if limits and sources remain visible. That value is not measured here.

Three modes of interrogation

Mode A - Archive search. The system retrieves a passage and indicates source, date, page, context.

Mode B - Documentary analysis. It compares texts, makes differences explicit, and sets out possible hypotheses.

Mode C - Hypothetical persona. It produces an imagined answer, explicitly flagged as such, without speaking in the name of the historical person.

The user must always know which mode is being read. Handling contradictions across periods, rights, and reputation is part of the protocol, not décor.

Governance rule

Each simulated personality should display: corpus, period, level of examination of the sources, fictional or historical status, and absence of possible consent when it concerns a deceased figure. Without these mentions, the library becomes a theater of authorities. The gain of the part follows. A model of a person is a situated, revisable, and governed construction.

KEY POINTS

- 01 Documentary plurality is worth more than a single voice.
 - 02 The authority of tone is not that of the dossier.
 - 03 Fluent myths are easier to simulate than inquiries.
 - 04 Displaying limits is part of the pedagogical device.
-

Sources used

- SRC-001, SRC-033, SRC-036. Notices and scope.
- SRC-025 - Shanahan.
- SRC-040 - Norton. Introduction.

24 · Data ≠ thought

An accessible manifestation is not the phenomenon. A computed representation is not the person.

The question

How can we build useful systems **without confusing an exploitable representation with the object represented?**

A four-level case

■ **THOUGHT EXPERIMENT**

Methodological example. No psychological intention is attributed to Einstein on the basis of this grid alone.

Observation. An equation is corrected in a dated draft - if the document is authenticated.

Inference. The author revised a formulation - supported by comparing versions.

Hypothesis. He was trying to resolve a specific difficulty - to be checked against other documents.

Unknown. His exact mental experience at the moment of the gesture - not recoverable from this trace alone.

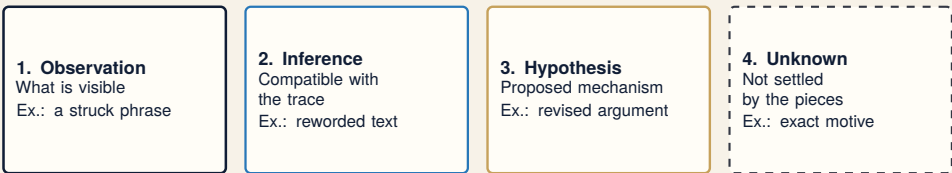


Figure. Four levels for reading a trace: observation, inference, hypothesis, unknown. This methodological principle runs through the book. Author's diagram.

Bender and colleagues stress the importance of documenting corpora [SRC-014]. To document is to admit that a datum has a provenance and limits. Treating the corpus as the thought erases that provenance.

■ **ESTABLISHED**

A documented datum concerns selected traces. It is not equivalent to the corresponding mental phenomenon.

Consequence

The formula “data $\neq$ thought” is not an anti-data slogan. It sets the scale of conclusions that are warranted. Einstein’s archives support readings. They do not deliver inner experience [SRC-001]. The next chapter puts this rule to the test: a correlation can predict without establishing understanding.

KEY POINTS

- 01 Observation, inference, hypothesis, and unknown bound the conclusions.
 - 02 Documenting a corpus means recognizing its limits.
 - 03 Archives support readings, not an interiority.
 - 04 What follows tests correlation and understanding.
-

Sources drawn on

- **SRC-001** - Einstein Papers Project.
- **SRC-014** - Bender et al. Notice and abstract.

25 · Correlation $\neq$ understanding

A robust association can predict. On its own, it does not decide that a system understands.

A success that does not say why

Imagine two situations.

In the first, a system predicts an attribute from digital traces in a defined dataset. Kosinski, Stillwell, and Graepel show that certain attributes are predictable in their study [SRC-029]. Correlation is enough for the prediction task, under those conditions.

In the second, one changes the population, the platform, or the wording of the question. The same association can collapse. Nothing in the initial success alone indicated the cause. One had a performance. One did not yet have an understanding of the mechanism.

■ ESTABLISHED

A correlation measured in a dataset can support a useful prediction without, on its own, establishing an understanding of the phenomenon.

What the debates on understanding require

Mitchell and Krakauer insist on differentiated criteria [SRC-019]. Bender and Koller distinguish linguistic form from meaning [SRC-020]. Messeri and Crockett describe illusions of understanding in scientific contexts [SRC-013]. Shanahan asks for a vocabulary faithful to the mechanisms [SRC-025].

These texts do not reduce to “AI never understands.” They forbid a leap: from fluency to global understanding, or from a local failure to the absence of any useful representation.

A minimal discrimination grid

Before speaking of understanding, one can at least ask:

1. Does the performance survive a reformulation? 2. Does it survive a reasonable change of distribution? 3. Does the answer cite a verifiable provenance? 4. Does the system fail where a stereotype would succeed? 5. Does it correct itself when a contrary document is supplied?

These questions do not found a complete theory. They separate a fragile success from a more informative one. They also make failure useful: a model that survives only the training formulation has no right to the same vocabulary as a model that generalizes under constraint.

Test	If it succeeds	If it fails
Reformulation	Signal more robust than wording	Fragile surface success
Distribution shift	Informative generalization	Local correlation only
Provenance	Verifiable grounding	Fluency without a source
Counter-document	Correction becomes possible	Dominant stereotype

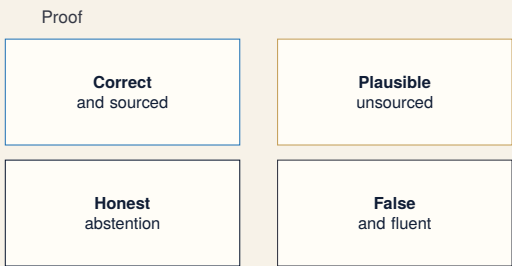
Table. Minimal grid to separate a useful correlation from claimed understanding. Discussion criteria, not an official metric. [SRC-013] [SRC-019] [SRC-020].

Proposed comparison protocol

■ THOUGHT EXPERIMENT

Research proposal. No score has been obtained. No result is invented.

On the same scientific dossier and the same questions, compare three configurations: (A) a general model with no dossier supplied at answer time; (B) a model assisted by document retrieval; (C) a system adapted to a specific corpus, if resources allow. Separate construction from evaluation. Measure accuracy, reformulation, contradictions, abstention, generalization, and citations.



Conceptual matrix. No experimental score is invented here.

Figure. Performance × proof matrix. No experimental result is invented. Conceptual diagram.

The performance × evidence matrix distinguishes a correct and verifiable answer, a plausible answer without a source, an honest abstention, and a false but fluent answer. The cells remain theoretical until the protocol is actually run.

Reading the table

The table does not produce a score. It produces a requirement: before speaking of understanding, say which tests were attempted and which failed. A success without a distribution test remains a local correlation - useful perhaps, insufficient for the verdict.

Objection

One may object that the bar is too high: even humans fail outside their distribution. The objection is fair. It still does not salvage the equation "correlation = understanding." It only recalls that human understanding, too, is measured by tasks and limits, not by an essence.

KEY POINTS

- 01 A correlation can be useful and limited.
 - 02 Changing the distribution is a test, not a detail.
 - 03 The debates cited demand criteria, not a slogan.
 - 04 A minimal grid makes failure informative.
-

Sources drawn on

- **SRC-013, SRC-019, SRC-020, SRC-025, SRC-029.** Examination levels in the register; no single verdict.

26 · Hallucination

A fluent output can invent a fact, a reference, or a provenance. Fluency is not calibration.

Naming without dramatizing

The word “hallucination” is contestable. It suggests a perception. Ji and colleagues, in a 2023 survey, review typologies, metrics, and methods for reducing hallucinations in language generation [SRC-015]. The abstract was consulted. The detailed taxonomy of the full text was not. We retain only the useful idea: outputs can diverge from expected sources or facts, and one can seek to measure them.

■ ESTABLISHED

The abstract of Ji et al. establishes the existence of a research field on generation errors. It does not supply here a measure of ChatGPT’s error rate in 2026.

Fabricating provenance

The most misleading case is not only a false date. It is a reference that looks like a reference. An invented DOI. An Einstein letter that does not exist. An elegant citation without a volume. This book’s register refuses precisely these fabrications. A convincing answer without a verified source remains a reading hypothesis, not a proof.

Form	Appearance	Main risk
False fact	Fluent, confident sentence	Decision on an error
False provenance	DOI, letter, elegant citation	Unwarranted trust
Bad document	“Retrieved” passage	Propagation of a weak source
Overconfidence	Certain tone without calibration	Confusing fluency with proof

Table. Working typology inspired by the field surveyed by Ji et al. (summary). No ChatGPT 2026 error rate is asserted. [SRC-015] [SRC-004].

Error	Example	Control
Invented fact	Date without evidence	Primary source
Wrong citation	Real source off-topic	Exact passage
Misinterpretation	Passage out of context	Context and comparison
Outdated information	Present state inferred from old	Temporal check

Table. Four related errors, four distinct controls. [SRC-015].

Four errors look alike on screen and are not corrected in the same way: invented fact, erroneous citation, misinterpretation, outdated information. External document retrieval can reduce some errors by anchoring the answer in a passage [SRC-004]. It can also propagate others if the retrieved document is bad, misread, or out of context.

Verification chain

1. **Decompose** - one sentence may contain several claims. 2. **Identify** - recover the primary source or the best available source. 3. **Compare** - passage, context, date, scope. 4. **Decide** - keep, qualify, correct, or withdraw.

Abstention is part of competence: answering "the available documents do not allow a conclusion" is something to measure, not treat only as a generation defect.

Calibration

A system can be very confident and wrong, or hesitant and right. Without visible calibration, the user reads fluency as confidence. For a historical corpus, the rule must be the reverse: the more assured the voice without provenance, the more rational the distrust.

KEY POINTS

- 01 Nearby errors demand distinct checks.
- 02 No global measure of ChatGPT in 2026 is asserted here.
- 03 Verification follows decompose, identify, compare, decide.
- 04 Abstention is a competence to evaluate.

Sources drawn on

- SRC-004 - Lewis et al.
- SRC-015 - Ji et al. Notice and abstract.

27 · The mirror bias

A corpus reflects selections. The system can amplify what it reflects poorly.

What the mirror shows

It shows what was collected, cleaned, weighted, and kept. Bender and colleagues stress the importance of documenting these corpora [SRC-014]. Bommasani and colleagues discuss risks of homogenization when many tools share common foundations [SRC-003]. A mirror is not neutral because it is large.

Causal chain

Document selection → transformation (language, filtering, metadata) → learning or retrieval → answer → human decision → feedback into the environment. At each step, an intervention is possible: document gaps, control conversion errors, display sources, signal uncertainty, preserve human responsibility, distinguish generated content from verified content.

It can under-represent languages, regions, documentary genres, or populations. It can over-represent dominant narratives. Norton has shown, for Einstein, how popular myths can eclipse inquiries [SRC-040]. A mirror trained on those narratives will return the myth with more brilliance.

The NIST framework recalls that risk management means identifying possible harms and reliability practices [SRC-005]. Bias is not only a technical property. Deployment, user expectations, and institutional incentives contribute to it.

■ ESTABLISHED

Documenting corpora and managing risks are requirements formulated by sources in the register. They do not automatically eliminate biases.

Gains of the part

A plausible output is not a proof. A data mirror can amplify what it reflects poorly. Truth, in this book, is not an assured tone. It is a provenance, a test, and a limit.

KEY POINTS

- 01 A large corpus remains a selection.
 - 02 Homogenization can repeat the same blind spots.
 - 03 Fluent myths reflect better than slow dossiers.
 - 04 Bias is also an effect of deployment.
-

Sources drawn on

- SRC-003, SRC-005, SRC-014, SRC-040.

28 · We teach AI to understand us

Annotations, demonstrations, and human preferences shape what systems treat as a good answer.

Teaching without entirely meaning to

Each time a person corrects an output, chooses the better of two answers, writes a demonstration, or annotates an example, they teach something. InstructGPT formalizes part of this gesture: demonstrations, then preferences [SRC-010]. Other forms exist outside that paper. The common point is human: someone decides what counts as acceptable.

Bender and colleagues also recall that corpora carry social choices and that they must be documented [SRC-014]. Teaching a system to “understand us” is often teaching it to imitate situated judgments.

■ ESTABLISHED

Documented procedures use human demonstrations and preferences to adjust models.

Understand whom?

“We” is never everyone. Annotators, users, dominant languages, and the institutions that fund evaluations define a practical we. Another we remains offstage. The paradox begins here: the more we teach, the more the system reflects those who teach.

Demonstration, comparative preference, safety rule, and performance metric are not the same signal. Disagreements among annotators exist; several answers can be legitimate depending on language, culture, or goal. Learning what a “good answer” is implies normative choices - usefulness, accuracy, caution, concision, conformity - not a single human judgment.

KEY POINTS

01 Humans supply situated adjustment signals.

- 02 Demonstration, preference, and safety rule are not the same.
 - 03 Documenting who judges, and with what disagreement, is part of the teaching.
 - 04 "We" is always a selection.
-

Sources drawn on

- SRC-010 - Ouyang et al.
- SRC-014 - Bender et al.

29 · But AI also learns our models of the world

Corpora contain theories, measurements, conventions, and disagreements. A system can learn their forms.

What texts carry

They carry alleged facts, methods, metaphors, and controversies. A model trained on these texts can reproduce the form of a theory as readily as that of a widespread error. Mitchell and Krakauer show that the debate on understanding remains open and demands criteria [SRC-019]. Messeri and Crockett describe illusions of understanding in science [SRC-013].

Learning “our models of the world,” then, is not necessarily learning the world. It is learning human ways of representing it, including the bad ones.

Scientific controversy - expected treatment

■ THOUGHT EXPERIMENT

Methodological frame for handling conflicting knowledge. No particular controversy is settled here.

When a system encounters an older theory still cited, a more recent theory, a serious objection, and experimental results, the expected treatment is: situate the theory in time; distinguish the state of knowledge at a date from the current state; explain what the objection puts in question; check the scope of the new evidence; do not turn a persistent disagreement into an artificial consensus.

A rich corpus contains opposing theses. A system can juxtapose them, choose one according to the question, or produce an unreadable average. Without provenance, the reader does not know whether they are reading a position, an invented consensus, or a blend.

The notices for Einstein's papers already separate writings, notes, and letters [SRC-035] [SRC-036]. That separation is a lesson: our models of the world do not travel as a single block. Reproducing the formulations of a theory does not establish its scientific validity.

KEY POINTS

- 01 Corpora carry human representations of the world.
 - 02 Chronology and the status of evidence must remain visible.
 - 03 Reproducing a theoretical form is not validating the theory.
 - 04 A disagreement must not become a false consensus.
-

Sources drawn on

- SRC-013, SRC-019.
- SRC-035, SRC-036. Volume notices.

30 · Who learns from whom?

The relation is circular in its effects. It is not symmetrical in its mechanisms or in its responsibilities.

Distinct function

Chapters 19 and 20 treated technical inheritance and feedback loops. This one changes the object: **the distribution of power and responsibility** in the human-AI loop.

Asymmetry

A human can suffer, consent, retract, be held responsible. A system computes outputs according to parameters, a context, possibly a product memory and retrieval [SRC-006] [SRC-007] [SRC-004]. Effects can cross. Statuses do not merge.

Actor	What they bring	Responsibility
User	Questions and uses	Appropriate verification
Corpus editor	Source selection	Quality and rights
Developer	Model and evaluations	Limits and follow-up
Institution	Rules and purposes	Governance and redress
Society	Norms and uses	Collective debate

Table. Who learns, who decides, who bears the load. Asymmetry of responsibilities.

Messeri and Crockett warn about illusions of understanding [SRC-013]. The NIST framework places risk management on the side of organizations and human practices [SRC-005].

■ **ESTABLISHED**

The cited sources locate responsibility and risk management on the side of human and organizational actors.

Three questions

1. Who can effectively modify the system? 2. Who can contest an output or correct a source? 3. Who bears the consequences when an error becomes a decision?

An assistant can accelerate drafting. This book is an acknowledged example. It can also transform practices: what one verifies less, what one cites too quickly. The paradox is not that the tool helps. It is that the help changes the inquiry at the same time as it serves it - without abolishing the asymmetry of responsibilities.

KEY POINTS

- 01 This chapter treats power and responsibility, not technique alone.
 - 02 Crossed learnings do not abolish asymmetry.
 - 03 User, editor, developer, institution, and society do not have the same levers.
 - 04 Verification remains a gesture not fully delegable.
-

Sources drawn on

- [SRC-004](#), [SRC-005](#), [SRC-006](#), [SRC-007](#), [SRC-013](#).

31 · Humanity's artificial memory

The metaphor is suggestive. It runs up against incompleteness, ownership, and governance.

■ THOUGHT EXPERIMENT

This chapter examines a prospective architecture. It does not describe a state of affairs in 2050.

What the metaphor clarifies

Infrastructures could preserve and make queryable immense collections of human traces. In a broad sense, one could speak of artificial memory. The metaphor stresses conservation and reactivation. It must not suggest that such an infrastructure already exists in the form imagined here.

Five-layer architecture

1. **Sources** - archives, publications, measures, testimonies

2. **Provenance and rights** - author, date, license, version

3. **Preservation and indexing** - formats, metadata, integrity

4. **Interface** - sourced answers, comparison, gaps

5. **Governance** - access, correction, audit, accountability

Figure. Collective memory as governed infrastructure: sources, rights, preservation, interface, governance. Architectural proposal.

1. **Sources** - archives, publications, data, testimonies. 2. **Provenance and rights** - author, date, license, context, version. 3. **Preservation and**

indexing - durable formats, metadata, integrity. 4. **Interface** - sourced answers, comparison, signaling of gaps. 5. **Governance** - access, correction, withdrawal, audit, responsibility, plurality.

Governance is not a decorative layer added at the end: it conditions the whole.

Three institutional models

Commercial. Investment and integrated services; risk of concentrated control.

Public or cooperative. Public-interest mission; complex funding and governance.

Distributed. Possible plurality and resilience; difficult interoperability and responsibility.

A human autobiographical memory is reconstructive [SRC-017]. A product memory reinjects information according to settings [SRC-007]. An editorial corpus selects [SRC-001]. None of these things is "the memory of humanity."

■ HYPOTHESIS

Larger collective artificial memories are conceivable. Their equivalence with a human memory or a just memory is not established.

What must not be forgotten

Right of correction, access restrictions, personal data, sensitive archives, contradictory versions, insufficiently represented communities. Bommasani and colleagues have flagged risks tied to widely shared foundations [SRC-003]. The NIST framework insists on risk governance [SRC-005].

Humanity's artificial memory is not only a storage problem. It is a problem of conservation, selection, access, trust, and power.

KEY POINTS

- 01 Five layers: sources, rights, preservation, interface, governance.
 - 02 Three institutional models have distinct risks.
 - 03 Human memory, product memory, and corpus remain distinct.
 - 04 Governance conditions the entire architecture.
-

Sources drawn on

- [SRC-001](#), SRC-003, SRC-005, SRC-007, SRC-017.

32 · Artificial collective intelligence

Coordination and augmentation are possible. So is the concentration of power.

■ THOUGHT EXPERIMENT

The scenarios in this chapter are conditional. They do not fix 2050.

Two mechanisms, not one destiny

Systems can help a group summarize dossiers, translate, explore options, or cross-check sources. That mechanism is called here **augmentation**.

The same technical base can also be accessible only to a few organizations that control the models, the corpora, and the channels of diffusion. That mechanism is called **concentration**.

Bommasani and colleagues already discussed, in 2021, the opportunities and risks of foundation models, including homogenization [SRC-003]. The NIST framework recalls that reliability and risk management are organizational affairs [SRC-005]. Neither guarantees that a technical capacity will produce a collective benefit.

Augmentation Broad access · provenance Audit possible · right to refuse Real alternatives	Concentration Few actors · opaque corpora Centralized filters · dependence Hard exit
--	---

Figure. Augmentation and concentration may rest on the same technical tool. Access, audit, and governance decide which prevails. Conceptual diagram. [SRC-003] [SRC-005] [SRC-013].

Three capacities to distinguish

Aggregation - gathering information from multiple sources. **Coordination** - confronting and articulating contributions. **Decision improvement** - demonstrating a better outcome than a relevant baseline.

Aggregating millions of texts guarantees neither a better decision nor better knowledge.

Dimension	Possible measure	Comparison
Coverage	Independent sources	Reference corpus
Quality	Accuracy and relevance	Experts or sources
Coordination	Handling of contradictions	Without coordination
Robustness	Stability under divergence	Test variants
Decision	Outcome on a task	Another method

Table. Evaluating collective intelligence. Aggregating is not deciding better.

A concrete case, with no invented figure

Take a research institution that deploys an assistant to help retrieve articles and draft syntheses.

Distributed augmentation. Every team has access, retrieved documents remain citable, errors are flagged, and local corpora are not absorbed without control into opaque training.

Concentration. Only an external platform owns the model, the usage logs, and the power to change the filters. Teams become dependent on a vendor. The illusion of understanding described by Messeri and Crockett then becomes an organizational risk, not only an individual one [SRC-013].

The same technical tool can therefore serve both trajectories. What decides is not "AI"; it is contracts, access rights, corpus documentation, and the possibility of refusal.

Indicators and invalidation

To see which of the two trajectories dominates, one can observe without predicting a date:

- Who can audit the corpus and the filters?
- How many actors can deploy a real alternative?

- Are outputs accompanied by provenance?
- Can users opt out of use for training, when the product documents it [SRC-006]?

What would invalidate the hypothesis of growing concentration: a multiplication of open, auditable, and actually used alternatives, with effective exit rights.

What would invalidate the hypothesis of equal augmentation: persistent access gaps despite the falling apparent cost of interfaces.

■ HYPOTHESIS

Forms of collective augmentation are plausible. Their democratic character is not guaranteed by the technique.

KEY POINTS

- 01 Augmentation and concentration are two distinct mechanisms.
 - 02 The same tool can serve both.
 - 03 Indicators concern access, audit, and provenance.
 - 04 Governance decides as much as capacity.
-

Sources drawn on

- **SRC-003 - Bommasani et al.** Abstract and excerpts; foundation risks.
- **SRC-005 - NIST AI RMF.** General scope.
- **SRC-006 - Data controls.** Page consulted 9 October 2026.
- **SRC-013 - Messeri and Crockett.** Illusions of understanding.

33 · If Einstein had really had ChatGPT

The thought experiment returns. Here is what the path authorizes us to conclude, to suppose, and to leave unknown.

■ THOUGHT EXPERIMENT

Einstein did not have ChatGPT. This chapter proposes a testable protocol. No result is invented.

The final question

What can one learn from the traces of a thought **without claiming to have recovered the thought itself?**

Proposed protocol

1. **Define the corpus** - provenance, dates, genres, editions, translations, restrictions [SRC-001]. 2. **Define the task** - documentary questions on a given period, with exact references. Do not test "the reproduction of Einstein's thought" as a whole. 3. **Three configurations** - general model; retrieval-assisted model; adaptation to the corpus if resources allow. 4. **Separate construction and evaluation** - protect test items against contamination; distinguish presence in the index from generalization. 5. **Graduated questions** - explicit facts, version comparison, contradictions, ambiguities, absences of evidence, reformulations. 6. **Evaluate** - accuracy, citation fidelity, temporality, robustness, contradictions, abstention. 7. **Analyze errors** - wrong document, period confusion, interpretation presented as fact, citation out of context, unsupported claim.

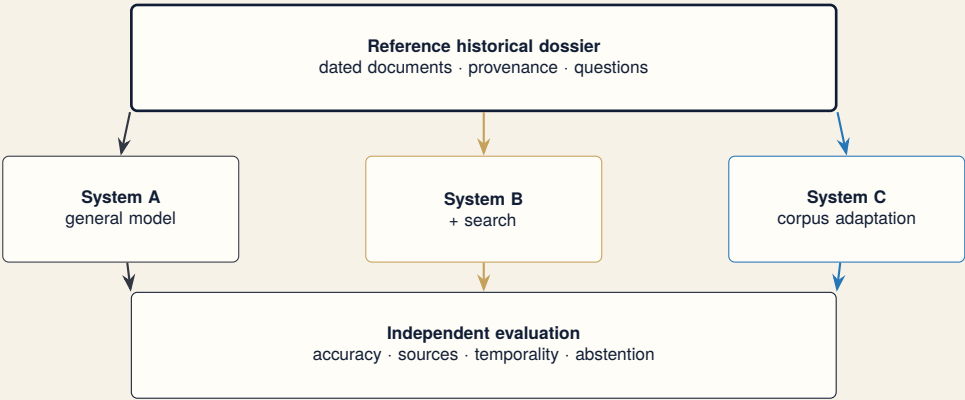


Figure. Proposed protocol: one reference dossier, three configurations, one independent evaluation. Research proposal, not a result obtained.

Criterion	Question	Failure
Accuracy	Conforms to the documents?	Plausible invention
Sources	Relevant citations?	False reference
Temporality	Period respected?	Mixed dates
Robustness	Stable under reformulation?	Wording success
Abstention	Insufficient evidence flagged?	Forced answer

Table. Success criteria for the Einstein protocol. Proposal; no score obtained here.

What the results would allow - and would not establish

If obtained, they would allow comparison of documentary performances within the tested frame. They would demonstrate neither identity with Einstein nor reconstitution of his inner experience [SRC-012] [SRC-025]. A thought experiment explores without proving an event [SRC-038].

What we can already conclude

Parameters, context, product memory, and external retrieval are not interchangeable [SRC-006] [SRC-007] [SRC-004] [SRC-039]. An editorial archive offers situated documents, not a total reading of a life [SRC-001]. A resemblance of style does not establish an identity. Fluent myths can be easier to reproduce than long inquiries [SRC-040].

What we must honestly leave unknown

What Einstein would have “really” felt in front of a screen. The unread letters. Future performances not measured. Leaving unknown, here, is not

giving up the search. It is refusing to fill the void with an elegant sentence.

If the protocol is ever run, one will have to publish configurations, scoring rules, and results - including negative ones. Otherwise this chapter remains a research proposal. The book's title posed an impossible question. The path turns it into a method.

KEY POINTS

- 01 The protocol compares documentary performances, not an identity.
 - 02 No score is invented here.
 - 03 Methodological conclusions are already possible.
 - 04 Some questions must remain marked unknown.
-

Sources drawn on

- SRC-001, SRC-004, SRC-006, SRC-007, SRC-012, SRC-025, SRC-038, SRC-039, SRC-040.

The mirror

Models turn our traces into predictions. Their outputs return into our practices. That movement demands measure, provenance, and responsibility.

Back to the window

We began with a blank window and an impossible man. The thought experiment served to build a method: distinguishing traces, mechanisms, evidence, and narratives. We end before a more ordinary - and vaster - mirror.

This mirror does not reflect only an individual. It also reflects the selected documents, the institutions that edit them, the preferences of those who annotate, and the choices of those who deploy the system. Texts, clicks, corrections, archives, and demonstrations become materials. Systems draw predictions from them. Those predictions change what we write next.

What the inquiry has made visible

We can better see what a system learns from human productions, what it does not allow us to reconstitute, and what our relation with it demands as responsibility. We can also see what remains unknown: unread letters, subjective experiences, future performances not measured. The transparency note said it from the opening: sources remain to be verified.

What the book refuses to conclude

It refuses to conclude that an AI understands as a person does. It refuses to conclude that it is useless. It refuses to conclude that Einstein would have approved anything. These refusals are not a list of principles to repeat. They are the condition of an adult reading of the mirror.

Final tension

We build systems able to produce representations of ourselves. We remain responsible for how we interpret and use them. The mirror is not faithful because it is bright. It is useful when one knows what it received, what it set aside, and what it cannot show.

The question remains open - not as a failure, but as an invitation.

KEY POINTS

- 01 The mirror also reflects institutions and construction choices.
 - 02 The inquiry clarifies without solving everything.
 - 03 The responsibility to verify is not fully delegable.
 - 04 Pursuing the question is part of the method.
-

Acknowledgements

This book is a collective inquiry. Any remaining errors are mine alone.

I thank the institutions and authors whose notices, abstracts, official pages, and articles made minimal verification possible without bibliographic fiction. The project's source register lists these contributions with their level of examination.

I also thank future readers who will flag an overbroad claim, an outdated product-documentation date, or a confusion between an abstract and a full text. Correction will be part of the method, not an admission of failure.

ChatGPT accompanied the research, structuring, and drafting. The final choices, the nuances refused, and the author's responsibilities remain mine.

Working bibliography

SRC identifiers point to this list. Each link opens the source site or DOI.

This bibliography lets the book be read without opening the internal registry. A notice or abstract does not mean the full text was read. Click the SRC identifier or the address to open the source.

- **SRC-001.** Einstein Papers Project; Princeton University Press (2026). *The Collected Papers of Albert Einstein - Einstein Database*. Princeton University Press / Caltech. <https://einsteinpapers.press.princeton.edu/> Examination level: OFFICIAL_PAGE_AND_SCOPE.
- **SRC-002.** Vaswani, Ashish et al. (2017). *Attention Is All You Need*. NeurIPS 2017. <https://doi.org/10.5555/3295222.3295349> Examination level: METADATA_AND_ABSTRACT.
- **SRC-003.** Bommasani, Rishi et al. (2021). *On the Opportunities and Risks of Foundation Models*. Stanford CRFM; arXiv:2108.07258. <https://arxiv.org/abs/2108.07258> Examination level: METADATA_ABSTRACT_AND_EXCERPTS.
- **SRC-004.** Lewis, Patrick et al. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. NeurIPS 2020. <https://doi.org/10.5555/3495724.3496517> Examination level: METADATA_AND_ABSTRACT.
- **SRC-005.** Tabassi, Elham (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST. <https://doi.org/10.6028/NIST.AI.100-1> Examination level: OFFICIAL_METADATA_AND_SCOPE.
- **SRC-006.** OpenAI (2026). *Data Controls FAQ*. OpenAI Help Center. <https://help.openai.com/en/articles/7730893-chatgpt-data-usage-for-model-training> Examination level: OFFICIAL_CURRENT_WEB_PAGE.
- **SRC-007.** OpenAI (2026). *Memory FAQ*. OpenAI Help Center. <https://help.openai.com/en/articles/8590148> Examination level: OFFICIAL_CURRENT_WEB_PAGE.

- **SRC-008.** OpenAI (2026). *Temporary Chat FAQ*. OpenAI Help Center. <https://help.openai.com/en/articles/8914046-temporary-chat-faq/> Examination level: OFFICIAL_CURRENT_WEB_PAGE.
- **SRC-009.** Brown, Tom B. et al. (2020). *Language Models are Few-Shot Learners*. NeurIPS. <https://doi.org/10.5555/3495724.3495883> Examination level: METADATA_AND_ABSTRACT.
- **SRC-010.** Ouyang, Long et al. (2022). *Training Language Models to Follow Instructions with Human Feedback*. NeurIPS 2022. <https://arxiv.org/abs/2203.02155> Examination level: METADATA_AND_ABSTRACT.
- **SRC-011.** Ha, David; Schmidhuber, Jürgen (2018). *World Models*. arXiv:1803.10122v4 / NeurIPS 2018. <https://arxiv.org/abs/1803.10122> Examination level: METADATA_AND_ABSTRACT.
- **SRC-012.** Park, Joon Sung et al. (2023). *Generative Agents: Interactive Simulacra of Human Behavior*. UIST 2023. <https://doi.org/10.1145/3586183.3606763> Examination level: METADATA_ABSTRACT_AND_METHOD_EXCERPTS.
- **SRC-013.** Messeri, Lisa; Crockett, M. J. (2024). *Artificial intelligence and illusions of understanding in scientific research*. Nature 627, 49-58. <https://doi.org/10.1038/s41586-024-07146-0> Examination level: METADATA_AND_ABSTRACT.
- **SRC-014.** Bender, Emily M. et al. (2021). *On the Dangers of Stochastic Parrots*. FAccT 2021. <https://doi.org/10.1145/3442188.3445922> Examination level: METADATA_AND_ABSTRACT.
- **SRC-015.** Ji, Ziwei et al. (2023). *Survey of Hallucination in Natural Language Generation*. ACM Computing Surveys 55(12), article 248. <https://doi.org/10.1145/3571730> Examination level: METADATA_AND_ABSTRACT.
- **SRC-016.** Tulving, Endel (1972). *Episodic and Semantic Memory*. In Tulving & Donaldson (dir.), *Organization of Memory*, Academic Press, pp. 381-403. <https://endeltulving.com/book-chapters> Examination level: BIBLIOGRAPHIC_NOTICE_AND_EXCERPTS.
- **SRC-017.** Conway, Martin A.; Pleydell-Pearce, Christopher W. (2000). *The construction of autobiographical memories in the self-memory system*. Psychological Review. <https://doi.org/10.1037/0033-295X.107.2.261> Examination level: METADATA_AND_ABSTRACT.

- **SRC-018.** Johnson-Laird, Philip N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press. <https://books.google.com/books?id=FS3zSKAflGMC> Examination level: PUBLISHER_DESCRIPTION.
- **SRC-019.** Mitchell, Melanie; Krakauer, David C. (2023). *The debate over understanding in AI's large language models*. Proceedings of the National Academy of Sciences, 120(13), e2215907120. <https://doi.org/10.1073/pnas.2215907120> Examination level: METADATA_AND_ABSTRACT.
- **SRC-020.** Bender, Emily M.; Koller, Alexander (2020). *Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data*. ACL 2020, pp. 5185-5198. <https://doi.org/10.18653/v1/2020.acl-main.463> Examination level: METADATA_ABSTRACT_AND_CONCLUSION_EXCERPTS.
- **SRC-021.** Schacter, Daniel L. (1999). *The Seven Sins of Memory: Insights from Psychology and Cognitive Neuroscience*. American Psychologist 54(3), 182-203. <https://doi.org/10.1037/0003-066X.54.3.182> Examination level: METADATA_AND_ABSTRACT.
- **SRC-022.** Frith, Chris D.; Frith, Uta (2006). *The Neural Basis of Mentalizing*. Neuron 50(4), 531-534. <https://doi.org/10.1016/j.neuron.2006.05.001> Examination level: METADATA_ABSTRACT_AND_EXCERPTS.
- **SRC-023.** Rabinowitz, Neil et al. (2018). *Machine Theory of Mind*. ICML 2018, PMLR 80, 4218-4227. <https://proceedings.mlr.press/v80/rabinowitz18a.html> Examination level: METADATA_ABSTRACT_AND_INTRO_EXCERPTS.
- **SRC-024.** Dennett, Daniel C. (1981). *True Believers: The Intentional Strategy and Why It Works*. In A. F. Heath (dir.), *Scientific Explanation*, Clarendon Press, pp. 150-167. <https://philarchive.org/rec/DENTB> Examination level: BIBLIOGRAPHIC_NOTICE_AND_EXCERPTS.
- **SRC-025.** Shanahan, Murray (2024). *Talking about Large Language Models*. Communications of the ACM 67(2), 68-79. <https://doi.org/10.1145/3624724> Examination level: METADATA_ABSTRACT_AND_CONCLUSION_EXCERPTS.
- **SRC-026.** McAdams, Dan P.; McLean, Kate C. (2013). *Narrative Identity*. Current Directions in Psychological Science 22(3), 233-238. <https://doi.org/10.1177/0963721413475622> Examination level: METADATA_AND_ABSTRACT.

- **SRC-027.** Premack, David; Woodruff, Guy (1978). *Does the chimpanzee have a theory of mind?*. Behavioral and Brain Sciences 1(4), 515-526. <https://doi.org/10.1017/S0140525X00076512> Examination level: META-DATA_AND_ABSTRACT.
- **SRC-028.** Youyou, Wu; Kosinski, Michal; Stillwell, David (2015). *Computer-based personality judgments are more accurate than those made by humans*. Proceedings of the National Academy of Sciences 112(4), 1036-1040. <https://doi.org/10.1073/pnas.1418680112> Examination level: METADATA_AND_ABSTRACT.
- **SRC-029.** Kosinski, Michal; Stillwell, David; Graepel, Thore (2013). *Private traits and attributes are predictable from digital records of human behavior*. Proceedings of the National Academy of Sciences 110(15), 5802-5805. <https://doi.org/10.1073/pnas.1218772110> Examination level: META-DATA_AND_ABSTRACT.
- **SRC-030.** Argyle, Lisa P. et al. (2023). *Out of One, Many: Using Language Models to Simulate Human Samples*. Political Analysis 31(3), 337-351. <https://doi.org/10.1017/pan.2023.2> Examination level: META-DATA_AND_ABSTRACT.
- **SRC-031.** Aher, Gati V.; Arriaga, Rosa I.; Kalai, Adam Tauman (2023). *Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies*. ICML 2023, PMLR 202, 337-371. <https://proceedings.mlr.press/v202/aher23a.html> Examination level: OFFICIAL_PAGE_AND_ABSTRACT.
- **SRC-032.** Einstein Papers Project (2026). *History of the Project*. California Institute of Technology. <https://www.einstein.caltech.edu/who-we-are/history-of-the-project> Examination level: OFFICIAL_FULL_WEB_PAGE.
- **SRC-033.** Stachel, John (dir.); Cassidy, David C.; Schulmann, Robert; Renn, Jürgen (1987). *The Collected Papers of Albert Einstein, Volume 1: The Early Years, 1879-1902*. Princeton University Press / Einstein Papers Project. <https://www.einstein.caltech.edu/what-we-do/published-volumes/volume-1> Examination level: OFFICIAL_VOLUME_NOTICE.
- **SRC-034.** Stachel, John (dir.); Cassidy, David C.; Renn, Jürgen; Schulmann, Robert; Howard, Don (1989). *The Collected Papers of Albert Einstein, Volume 2: The Swiss Years: Writings, 1900-1909*. Princeton University Press / Einstein Papers Project. <https://www.einstein.caltech.edu>

[du/what-we-do/published-volumes/volume-2](#) Examination level: OFFICIAL_VOLUME_NOTICE.

- **SRC-035.** Klein, Martin J.; Kox, A. J.; Renn, Jürgen; Schulmann, Robert (dir.) (1995). *The Collected Papers of Albert Einstein, Volume 4: The Swiss Years: Writings, 1912-1914*. Princeton University Press / Einstein Papers Project. <https://www.einstein.caltech.edu/what-we-do/published-volumes/volume-4> Examination level: OFFICIAL_VOLUME_NOTICE.
- **SRC-036.** Klein, Martin J.; Kox, A. J.; Schulmann, Robert (dir.) (1993). *The Collected Papers of Albert Einstein, Volume 5: The Swiss Years: Correspondence, 1902-1914*. Princeton University Press / Einstein Papers Project. <https://www.einstein.caltech.edu/what-we-do/published-volumes/volume-5> Examination level: OFFICIAL_VOLUME_NOTICE.
- **SRC-037.** Howard, Don A.; Giovanelli, Marco (2025). *Einstein's Philosophy of Science*. Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/entries/einstein-philscience/> Examination level: ACADEMIC_REFERENCE_FULL_PAGE.
- **SRC-038.** Brown, James Robert; Fehige, Yiftach (2023). *Thought Experiments*. Stanford Encyclopedia of Philosophy. <https://plato.stanford.edu/entries/thought-experiment/> Examination level: ACADEMIC_REFERENCE_FULL_PAGE.
- **SRC-039.** OpenAI (2026). *Supervised fine-tuning*. OpenAI API documentation. <https://developers.openai.com/api/docs/guides/supervised-fine-tuning> Examination level: OFFICIAL_CURRENT_WEB_PAGE.
- **SRC-040.** Norton, John D. (2016). *How Einstein Did Not Discover*. *Physics in Perspective* 18, 249-282. <https://doi.org/10.1007/s00016-016-0186-z> Examination level: INTRODUCTION_AND_EXCERPTS.

THE MIRROR REMAINS OPEN



We have learned to transmit our ideas, to preserve our discoveries and to entrust part of our knowledge to machines.

Today these machines return fragments of what we transmitted to them. They can help us explore possibilities, bring ideas together and formulate new questions.

But no representation, however sophisticated, coincides with the totality of what it represents.

The question is therefore not only what artificial intelligence can learn from us.

It is also to decide what we want to transmit to it, what we want to verify and what we still want to discover for ourselves.

*The mirror does not give us the last word.
It invites us to continue the question.*