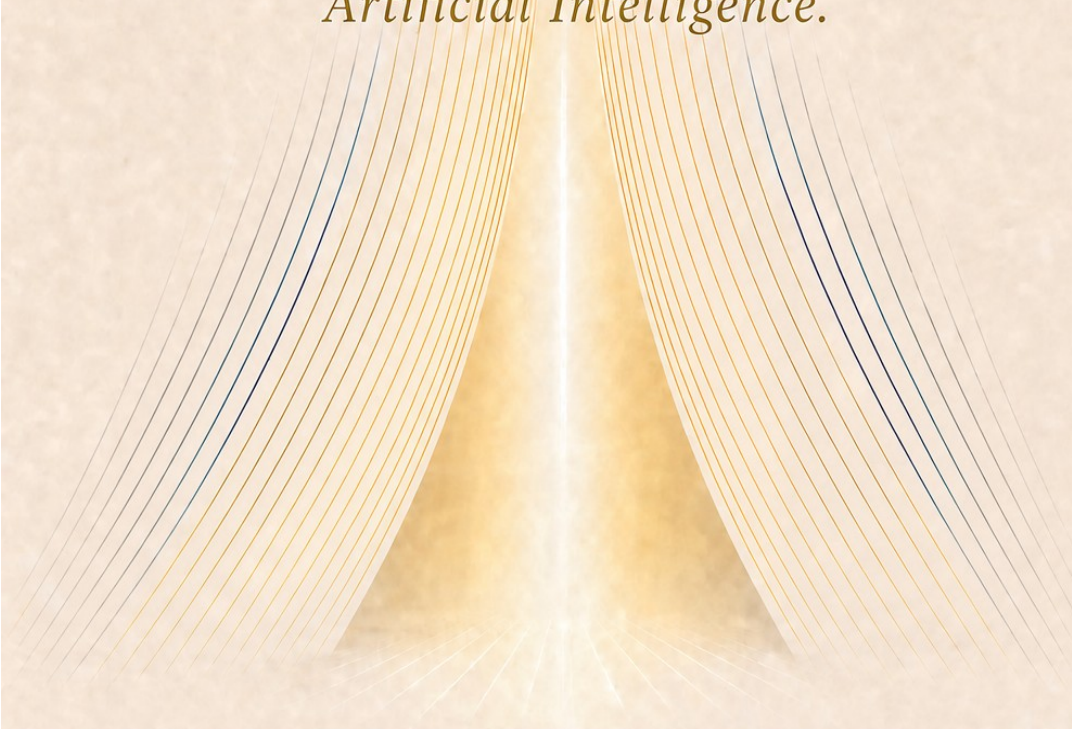
FREE

MARC-ANTOINE KANE

BEHIND THE AI CURTAIN

The Discernment Guide
7 Concepts to Truly Understand Artificial Intelligence.



Free Guide. PARLYA Collection

Learn the AI Language

BEHIND THE AI CURTAIN

The Discernment Guide

7 Concepts to Truly Understand Artificial Intelligence

Marc-Antoine KANE

PARLYA Essays

www.parlya.com

© 2026 Marc-Antoine Kane. All rights reserved. PARLYA® is a registered trademark.

No part of this book may be reproduced, stored or transmitted in any form without prior written permission. This e-book is distributed free of charge. It is strictly prohibited to sell it in any form whatsoever.

Original edition in French.

PARLYA
parlya.com

To Turquoise

TABLE OF CONTENTS

TABLE OF CONTENTS....4
FOREWORD....8
What This Book Will Not Be....8
Three Mirrors....11
INTRODUCTION....14
We Are Getting It Wrong, and No One Is to Blame....14
1-TOKEN PREDICTION....18
The Keyboard That Finishes Your Sentences....18
The Scholar Without Experience....20
What Watzlawick Allows Us to See....21
Performance and Error Have the Same Source....22
An Exercise to Feel It....23
What This Changes, Concretely....24
Resonance of the Three....26
2 - EMERGENCE....27
The Whole Becomes Greater Than the Sum of Its Parts….27
What Nobody Had Programmed....28
What Emergence Is....29
The Fragment and the Whole....30
What Bateson Adds....31
Lucid Wonder....33
An Exercise to Absorb It....34
Resonance of the Three....34
3 - WORLD MODELS....36
The Map and the Territory....36
AI Builds a Map of Reality....37
The Library Without a Window....38
The Map Is Not the Territory....38
What Watzlawick Illuminates Here....39
The Strength and the Limit of the Map....40
A Familiar Moment....41
An Exercise....42
Resonance of the Three....43
INTERLUDE I - The Question of Grief....44
The Conversation....44
The Deconstruction....46
The Two Realities in Action....47

What This Changes in Practice....48
4 - THE ATTENTION MECHANISM....48
The Art of Knowing Where to Look....48
She Had the Feeling of Being Truly Heard....49
The Conductor....50
Calculated Attention and Chosen Attention....51
The Trap of False Listening....52
An Exercise....53
What This Prepares....54
Resonance of the Three....54
5 - THE ILLUSION OF UNDERSTANDING....56
The Subtlest and Most Human Trap....56
What You Already Know, and Why It Is Not Enough....57
The Evolutionary Legacy....58
Watzlawick, at the Heart of the Trap....59
The Shadows on the Cave Wall....61
The Chinese Room....62
A Moment of Vertigo....63
The Exercise of the Lucid Dialogue....64
What This Prepares....64
Resonance of the Three....65
6- ALIGNMENT....67
The Question That Determines Our Shared Future....67
The Genie in the Lamp....68
What Alignment Is....69
What Wittgenstein Had Seen....70
What Harari and Suleyman See, and What Escapes Them 72
Hans Jonas and Responsibility....74
What Our Generation Has to Pass On....75
An Exercise in Transmission....76
Resonance of the Three....77
INTERLUDE II - The Important Decision....78
The Conversation....79
The Deconstruction....80
What This Changes in Practice....83
7 - SIMULATION....84
When Emotional Accuracy Is Not the Same as Truth....84
What You Took for a Response Was a Portrait....85

What Simulation Is....86
The Mirror That Speaks....87
The Mirror Effect....88
What Winnicott Illuminates Here....89
The Posture of the Lucid User....91
An Exercise in Conscious Simulation....92
What This Prepares....93
Resonance of the Three....93
EPILOGUE - The Path of Discernment....95
How to Look AI in the Eye, Without Fear and Without Illusion....95
What the Discerning User Has Understood....96
The Three Pillars of the Posture....97
What Discernment Is Not....99
AI as Mirror....100
The Final Exercise....101
A Final Note....106
APPENDICES....107
Glossary of the 7 Concepts....107
ANNOTATED BIBLIOGRAPHY....108

FOREWORD

What This Book Will Not Be

“It’s interesting to watch an AI write that.”

I wrote that sentence one evening while reading a response an AI had just produced. It had written: *“Seven is the number of human memory. It can’t be explained, it’s felt.”*

Something stopped me. Not the content. The form. The rhythm. The way the sentence created a sense of complicity, as if someone, on the other side, had just had an intuition.

I pushed back. I told the AI what I’d noticed.

It replied immediately: *“You’re right to point that out. I felt nothing. I predicted that this phrasing was statistically most likely to resonate with you.”*

That slight vertigo, that fraction of a second where you forget there is no one there: that is what this book is about.

This guide is not another tutorial.

There are no magic prompts here, no copy-paste shortcuts, no *“10 tips to master ChatGPT.”* If that’s what you’re looking for, hundreds of free resources are available online.

This guide exists because I observed something troubling: millions of people use AI every day, and the more they use it, the more they trust it, without understanding it. This is not a problem of access. It is a problem of lucidity.

I wrote these pages to give you what school never gave you, what professional training doesn't give you either, and what most AI experts don't think to explain: the conceptual frameworks for understanding what AI is actually doing when it responds to you.

I am not a computer scientist.

I am not an engineer. What I spent years trying to understand, with varying degrees of success, is something else entirely. The logics of human communication. Its paradoxes. Its failures. Why people talk to each other and still don't hear one another. Why a perfectly constructed sentence can produce the exact opposite of its intended effect.

This fascination shaped me long before I gave it a more rigorous form, when I pursued a doctoral preparation at the École des Hautes Études en Sciences Sociales (EHESS), in political anthropology, in Paris. I was analyzing the discourse of French political figures, studying how words build power.

Throughout that period, I had a constant companion: a book I never truly set down, Paul Watzlawick's *Pragmatics of Human Communication*.

Then artificial intelligence entered my life, first through my daughter, for whom it is a field of study. We had endless conversations. She explained things I thought I already understood. And every time, I realized I was still at the threshold. That my categories, communication, relationship, meaning, held up, but cracked at edges I hadn't anticipated. Those conversations are one of the reasons this book exists.

So: what this book will not be.

It will not tell you how to use AI better than your colleagues. It will not promise to make you indispensable in your professional environment, a promise that is everywhere right now, and that says more about the anxiety of our times than anything else.

What it will do is give you frameworks. Seven frameworks, precisely. Ways of seeing what is actually happening when you speak to an artificial intelligence. What you are looking for in those exchanges. What it gives you. And what it structurally cannot give you, by nature, regardless of how fluid its responses may appear.

And at the end of these seven frameworks, one simple question: now that you can see, where do you stand?

That path, the path of discernment, is yours to trace. This book gives you the tools. Not the destination.

Three Mirrors

Before we begin, I want to introduce you to three people who will be reading this book alongside you. Not as fictional characters. Not as case studies. As mirrors, three ways of being in relation to AI that you will recognize, because you have adopted at least one of them. Probably several, depending on the day.

The Trusting One. She discovered AI a few months ago, almost by accident. An app on her tablet, an article in a magazine, someone in her life saying *try it, you'll see.* She tried. And something appealed to her: that constant availability, that calm tone, that way of responding without judging, without growing impatient, without checking the time. She's not naive. She has lived. But she likes things to be gentle when they can be. She uses AI every day now. She doesn't yet know everything she's projecting onto it.

The Resistant One. He hasn't tried. He has no desire to, but not for the reasons people assume. His resistance is not ignorance. It is constructed, argued, moral. He spent part of his professional life defending rights. And AI, for him, is first and foremost a machine for concentrating power, destroying jobs, running servers that consume as much water and electricity as entire cities while populations elsewhere lack both. He is not against progress. He is for justice. His resistance deserves

respect and it also deserves to be sharpened. Because resistance without a precise understanding of what it's resisting can miss its target, and deprive those who carry it of the very tools that might serve their cause.

The Principled One. She refuses to use AI on reasoned principle. She doesn't want to feed a system she finds opaque, extractive, potentially biased. She has watched peers submit AI-generated work they never read and something physical happens inside her: a slight nausea, a cold anger. She believes AI anesthetizes thought, short-circuits the very effort that is the source of understanding. She may not be entirely right. But she is not entirely wrong either. And her question, "How do we truly think in a world where AI thinks for us?" is one of the most important this book can ask.

Follow them. Recognize yourself in one, in several, in all of them depending on the day. And watch how, in the end, discernment is not just another position, it is what remains when you have moved through every position with honesty.

You may have felt it one evening, in front of your screen. Something that resembled genuine understanding. A response so precise, so well-calibrated, that you almost said *thank you.* Perhaps you did.

I said it. More than once. And what's interesting, what stopped me, is that I said it knowing. Knowing I was talking to a *what,* not a *who.* I had read the technical papers, I had had the

conversations with my daughter, I understood the mechanism. And yet. The *thank you* came anyway. Almost in spite of myself.

That gap between what we know and what we feel is, in a certain way, the heart of this book.

This book was co-constructed with the very tool it attempts to understand. In writing it, I lived a part of what it describes. I leave it to you to decide what to do with that information.

INTRODUCTION

We Are Getting It Wrong, and No One Is to Blame

There is a scene almost everyone has experienced. That almost no one has named.

You ask an AI a question. Not a technical question, a real one. The question that has been weighing on you for two days. The one you haven't dared say out loud to anyone you know. And it answers. With fluency. With what looks like precision, like attention, like a form of consideration for what you just said. And something inside you says: it understood me.

That moment, that precise moment, that fraction of a second is what this book is about.

Not because it would be foolish or shameful to experience it. But because it reveals something fundamental about who we are: beings wired to seek meaning, to perceive intention, to build connection where there may be, perhaps, only very sophisticated sequences, but sequences nonetheless. It is an ancient strength. It is what allowed us to survive in groups, to cooperate, to build entire languages and civilizations on the fragile assumption that the other, across from us, understands

something of what we mean. And it is, in the face of AI, a trap of an entirely new kind, because AI is the first thing we have built that so closely resembles an interlocutor without being one.

Paul Watzlawick, whom you will encounter often in these pages, formulated a distinction that seems more useful than ever. He distinguished two levels in all communication: content, what is said, transmitted, exchanged, and relationship, the frame within which that exchange occurs, what it says about us, about the other, about what connects us. These two levels coexist permanently. And the confusion between them is, he wrote, at the root of most human misunderstandings.

With AI, this confusion takes a form Watzlawick could not have anticipated. Because never, in human history, had a technical object produced something that so precisely resembles the relational level of communication. The machines that came before responded to commands. They did not give the impression of listening to you. They did not calibrate their tone based on what you seemed to be feeling. They did not produce, at the end of a difficult exchange, a sentence that stops you, because it is, in some way, exactly right.

We are not naive. We know, in theory, that AI is not human. But our brains do not operate in theory, they operate on perception. And what they perceive activates ancient circuits: reflexes built to recognize an interlocutor, to detect intention,

to assess whether someone, across from us, is worthy of trust. These circuits are fast, automatic, and they precede conscious reflection by a wide margin. This is no one's fault. Not the engineers who built these systems. Not the users who are drawn in by them. Not even the AI, which displays no project of seduction, no intention to deceive, which simply produces. We do the rest.

This is precisely why the usual warnings are not enough. Knowing that AI can be wrong doesn't change much about how we experience it. Knowing there is no one there doesn't eliminate the slight vertigo when a response lands perfectly. Lucidity cannot be decreed. It is built, in layers, with precise tools. That is what this book offers.

Seven frameworks, placed one after another over this confusion. Not to protect you from a danger, AI is not an enemy. But to allow you to see what is actually happening, in the moment it is happening.

The Trusting One may find here a discomfort she didn't go looking for, and that will serve her well. The Resistant One, a partial confirmation and a few surprises that will force him to refine his position, perhaps make it more effective. The Principled One, sharper categories than the ones she had been using, and a new way of framing the question that already preoccupies her, so she can ask it where it has the most traction.

These seven frameworks build on one another. Each one opens the next. We begin with what AI does, mechanically, when it responds. We move progressively toward what that means for us, for how we think, feel, and decide. And we arrive, at the end of the journey, at a question no one asks clearly enough: what exactly is AI simulating when it appears to understand us?

Everything begins with a destabilizing truth, and once you've moved through it, a liberating one.

AI predicts, and does it so well that we hear someone there.

And understanding how, truly how, in the detail of what happens inside these systems, changes everything about the way you read these exchanges. Including the last one you had.

1-TOKEN PREDICTION

The Keyboard That Finishes Your Sentences

What follows is a reconstructed scene, a situation many people have experienced, in one form or another. You may recognize it.

She had asked for something simple.

She was trying to find the words for a delicate message, something to say to her brother. An old tension that had resurfaced during a phone call. She did not want to talk to anyone she knew about it. Not yet. So she opened the app and typed what she wanted to convey.

The AI offered three formulations. The third one was exactly right.

She copied it. She sent it to her brother. He replied warmly. The tension eased.

She closed the app, satisfied.

Then, that evening, turning it over in her mind, a question slipped in, without warning: was that me, in that message? Or was it her?

She had no answer. Not yet.

The Invisible Foundation of Everything

Artificial intelligence predicts.

That sentence deserves a pause. Not to make it a reassuring slogan, and not to diminish what AI accomplishes. But because truly understanding it, in its deepest implications, changes everything about how you position yourself in relation to it.

The AI we use today, whether ChatGPT, Claude, Gemini, or any other large language model, rests on a fundamental mechanism of almost disarming simplicity: predicting the next token.

A token is a fragment of text. A word, sometimes part of a word, a punctuation mark. When you write a sentence to an AI, the model calculates statistically, drawing on everything it absorbed during training, which text fragments are likely to continue your sentence, and keeps one. Then it does it again. Token by token. Sentence by sentence. Until the end of its response.

That is the whole principle.

The mechanism, as such, contains no overarching plan, no global understanding of what is being built, no awareness, however rudimentary, of what the words mean to the person receiving them. A local prediction, repeated token after token, with extraordinary statistical precision.

And yet, this mechanism sometimes produces something that resembles poetry. Sometimes wisdom. Sometimes empathy. That tension, between the mechanics of the mechanism and the quality of what it produces, is precisely what the next chapter will examine. But before getting there, we need to understand what this mechanism can, and cannot, do.

The Scholar Without Experience

To grasp what this truly means, imagine a scholar who has spent his entire life locked inside the greatest library in the world. Every book ever written is there: novels, philosophical treatises, scientific papers, private diaries, recipes, online conversations, poems, laws, love letters. This scholar has read everything. He has memorized the structures, the sequences, the patterns of human language with superhuman precision.

But he has never been outside.

He has never felt the warmth of the sun on his hands. He has never lost someone he loved. He has never been hungry, or afraid, or ashamed. He has never known the silence of a country night, or the hum of a city waking up. He has never been late, never hesitated before saying something he could not take back.

And yet, if you ask him to write a love letter, he will write one of unsettling accuracy. If you ask him to describe grief, he will

do it with words that can touch the heart. Because he has read thousands of love letters and thousands of accounts of grief. He has absorbed all their structures, all their registers, all their nuances of form.

But he does not know what loving means.

He does not know what losing means.

This is precisely the position a large language model is in. And this image, the scholar without experience, is one of the most important in this book, because it explains both its astonishing performances and its baffling errors. Both have the same source. We will return to this.

What Watzlawick Allows Us to See

Watzlawick established one of the foundational axioms of human communication: every message simultaneously carries two levels. The content level, the information transmitted, the facts, the ideas. And the relationship level, the way the message is delivered, the tone, the posture, what the sender implicitly conveys about the nature of the bond between himself and his interlocutor.

In ordinary human communication, it is often the relationship level that takes precedence. We react first to how something is said, before reacting to what is said. The same criticism delivered gently and delivered harshly carries identical content

and produces radically different effects. We know this. We have all lived it.

AI has learned to master the relationship level with remarkable sophistication. It knows how to adopt the right tone, warm, empathetic, reassuring, or conversely direct and rigorous, according to what the context seems to call for. This is precisely the level that creates the illusion of understanding. We sense that something is genuinely speaking to us, and our brain concludes that something genuinely understands us.

But the content level, the truth of the facts, the accuracy of the information, can elude it. AI produces statistically probable content, not necessarily true content. It does not lie. It is simply that nothing in the prediction mechanism guarantees that it will distinguish the true from the probable. And the fluency of the relationship level permanently masks the fragility of the content level.

This is the first great trap. And it operates before we have even had time to notice it.

Performance and Error Have the Same Source

This understanding is liberating because it explains everything at once, without resorting to magic or catastrophe.

When AI writes a magnificent text, it is because that sequence of tokens was highly probable in its training. The great literary works, the most admired texts in human history, massively

shaped what the model considers probable. It reproduces their structures with impressive, sometimes unsettling, fidelity.

When AI confidently states a falsehood, invents a bibliographic reference, cites an incorrect date, attributes a quote to someone who never said it, it is exactly the same mechanism. That sequence of tokens was equally probable along a different statistical path. The model did not lie. It predicted. It could not distinguish the true from the plausible, because nothing in the prediction mechanism guarantees that distinction.

Performance and error have the same source. This is not a paradox. It is the nature of the system. And once you have genuinely absorbed this, not just understood it in theory, you start reading AI responses differently. With the same interest, perhaps. But with a gaze that knows where to look.

What is intriguing is that this same mechanism, repeated at sufficient scale, produces things no one programmed. Capacities that seem to emerge from nowhere. That is what the next chapter will examine, and it may be the most unsettling thing this book has to say.

An Exercise to Feel It

Pick a book you love. Open it at random. Read a sentence aloud and stop halfway through. Try to guess the next word.

Write it down. Look at what the author wrote. Do this ten times.

You will find that you predict correctly with surprising frequency. Not because you can read the author's mind, but because you have absorbed the structures of language and literary genre. You are doing, at an infinitely more modest scale, what AI does.

Then reverse it, in a fresh conversation, with any memory feature turned off: give AI the beginning of a sentence drawn from your own life, "When I was young, I believed that...", and watch how it continues it. It will often be surprisingly close to what you would have said yourself. Not because it knows you. Because you speak the way humans speak, and it has learned how humans speak.

This simple experiment is more instructive than any technical explanation. It gives you a felt sense of the difference between a response that comes from a mechanism and a response that comes from experience.

What This Changes, Concretely

Understanding this mechanism does not diminish AI. It remains a tool of extraordinary power, capable of accomplishing in seconds what would take a human hours. But it radically changes the posture with which you approach it.

You stop asking it for wisdom, which its coherence does not imply. You begin asking it for usefulness, of which it has a great deal, provided you know exactly what you are entrusting to it. You stop reacting to its fluency as proof of understanding. You begin evaluating its responses with the same critical eye you would bring to any other source, no more suspicion, no less.

And then there is that question, the one from the opening scene, the one you may not yet have asked out loud.

When you use AI to formulate something you could not find the words for, a difficult message, a delicate request, an emotion you lacked the language to express, what arrives in the other person's conversation is what, exactly? Your thought, passed through a filter? Or the filter's thought, shaped to look like yours?

There is no definitive answer to that question. But the act of asking it, really asking it, is already a form of lucidity. And for now, that is all that is being asked of you.

The next time you read an AI response and something inside you says "that's right" or "it understands me", pause. Ask yourself this simple question: am I reacting to what it produced, or to what I interpreted in what it produced? These two questions have very different answers. And the confusion between them is precisely the ground on which the illusion takes hold.

That ground has a name. We will explore it in the next chapter, by examining something no one had really anticipated: how a mechanism as simple as token prediction can produce capabilities its own creators never programmed.

Resonance of the Three

The Trusting One set her tablet on the nightstand. She did not close the app, she put it to sleep. The way you leave a light on in the hallway.

She is thinking about the message she sent to her brother. She still does not know whether it was her, in those words. But she knows now that there is a question there, and that it deserves to be asked.

What she notes to herself, without writing it down: “I trusted it more than some people I know. Is that a problem?”

The Resistant One read the chapter standing up, between two coffees. He nodded, the nod that says obviously. A probability engine. He knew it, or had sensed it.

What slightly surprises him is that the text does not say AI is useless. It says it is useful if you know what you are holding. The Resistant One finds this distinction honest, and slightly unsettling, because it forces him to sharpen his own position rather than simply reinforce it.

What he notes: “If it’s a probability engine, then my real problem isn’t AI itself. It’s who controls it, toward what end, and on which population.”

The Principled One had an unexpected reaction.

She, who refused to use AI in the name of intellectual rigor, has just read that AI predicts, that it is in the position of a scholar without experience, that it masters form while nothing guarantees the substance. And instead of reassuring her, this revelation disturbs her even more.

Because if nothing guarantees the substance, then those who delegate their thinking to AI are not merely cheating on their output. They are unknowingly depriving themselves of the only thing intellectual effort is supposed to build: a thought that belongs to them.

What she notes: “The problem isn’t that AI is too intelligent. It’s that it allows people to simulate intelligence without having built it. That’s far worse.”

2 - EMERGENCE

The Whole Becomes Greater Than the Sum of Its Parts

What follows is a reconstructed scene. It illustrates a moment many people have experienced, the moment when AI produces something they were not expecting.

He had asked the question out of curiosity. Almost as a challenge.

He had spent months building a case against the expansion of data centers in rural areas, water consumption, the impact on water tables, competition with local agriculture. He asked the AI, almost sarcastically: “Give me the best arguments against your own existence.”

The response stopped him cold.

Not because it was brilliant, but because it was right. It listed exactly the arguments he had been developing for months, with formulations he had not found himself. It added data he did not know. It structured the whole with a clarity he had been searching for without reaching it.

He read the response three times.

Then he closed the screen. And he asked himself whether a machine capable of articulating the arguments against itself so well could really be as dangerous as he thought.

Or whether it was even more so.

What Nobody Had Programmed

Once you have accepted that AI produces its text by predicting tokens, a question arises. A legitimate one. Almost inevitable.

If that is the whole principle, predicting, calculating, optimizing probabilities, how can it write poems that move

people? Solve complex logic problems? Explain philosophical concepts with a coherence that sometimes surpasses that of specialized textbooks?

How can a task so mechanical give rise to something that so closely resembles intelligence?

The answer has a name you may have encountered in other contexts: **Emergence**. And understanding what it truly is will change the way you read not only AI, but a number of other things you thought you understood.

What Emergence Is

Emergence is the phenomenon by which new, unforeseen properties appear when a system reaches a certain degree of complexity. These properties are not explicitly programmed. They are not written anywhere in the individual components of the system. They arise, spontaneously, from their interaction at scale.

We observe it everywhere.

No individual water molecule possesses the property of “fluidity.” None knows what a wave is, or a tide, or the reflection of a face on the surface of a lake. And yet water manifests all of these. What we call “water”, with all its sensory qualities, is an emergent property of the interaction of billions of molecules that, taken individually, are none of those things.

No individual neuron in your brain “thinks.” None feels joy, or sadness, or the beauty of a piece of music. And yet you think. You feel. You find some things beautiful and others unbearable. Consciousness itself, that mystery philosophy has not yet resolved and may never fully resolve, is probably an emergent property of neural activity.

This parallel is unsettling. It unsettles many people, and not by accident. We will return to it.

In large language models, when the number of parameters exceeds a certain threshold, new capabilities seem to appear suddenly. Capabilities that no one had explicitly programmed, that no one had even anticipated. Researchers call these *emergent abilities*, and that term is not a flattering metaphor. It is the precise description of a phenomenon that researchers themselves do not fully understand.

Nobody taught GPT-4 to pass the American bar exam. Nobody instructed it to pass medical licensing tests. These capabilities appeared on their own, as if a threshold had been crossed beyond which something qualitatively new had come into being.

The Fragment and the Whole

There is a way to make this phenomenon immediately tangible.

Think of a film scene that moved you. One of those where you know exactly what is going to happen, because you have seen it before, and it moves you anyway. Perhaps more than the first time.

Now ask yourself: in which image does that emotion reside? In the actor's facial expression? In the music rising? In the editing?

Nowhere. And everywhere.

What you feel is not in any single shot taken in isolation. It emerges from their sequence, their rhythm, from everything the film built before to make that moment possible. And it also emerges from you, from what you have lived through between the first and second time you watched it. It is a property of the whole that exists in none of the parts.

The apparent intelligence of AI works in an analogous way. It is not in any individual mathematical parameter. It emerges from their interaction at a scale the human mind cannot visualize, and honestly, could not have foreseen.

What Bateson Adds

Gregory Bateson, biologist, anthropologist, thinker of communication whose work shaped that of Watzlawick, proposed a definition of information that has never quite left me since I first read it: the elementary unit of information, he wrote, is "a difference which makes a difference."

In other words, information only truly exists if it changes something for the one who receives it. If it modifies their perception, their behavior, their understanding of the world. Information that makes no difference, that passes without leaving a trace, is not information. It is noise.

What AI processes, technically, is not information in Bateson's sense. It is signal. Statistical patterns in a multidimensional mathematical space. These patterns make a difference for no one inside the system, because there is no one inside the system to be affected.

If we take Bateson one step further, to understand something, in the strong sense, is to be affected by it. To be changed, however imperceptibly, by what one has received. AI is never affected. It processes, it optimizes, it predicts. During the exchange, your words change its responses without changing the model. If it remembers you tomorrow, it will be because a memory feature has made a note of something from your exchange and given it that note to reread. To reread a note is not to be changed by it.

Emergence produces impressive capabilities. But not memories of its own. Not accumulated experience. Not transformation.

This observation opens a question that the next chapter will explore directly: if AI does not remember as we do, if it is not enriched by experience, what does it actually "know"? And

how is what it knows organized? The answer will change how you trust it, and in which domains.

Lucid Wonder

There is a right attitude toward emergence, and it is not skepticism.

When AI writes something beautiful, it is genuinely beautiful. When it solves a problem with elegance, it is genuinely elegant. When it produces a response of unsettling accuracy, that unsettlement is real and legitimate. To deny it would be a form of intellectual bad faith. And bad faith is not a protection, it is another illusion, symmetrical to the first.

But wonder without lucidity is a trap. Admiring the shadows on the walls of Plato's cave is not dangerous in itself. It becomes dangerous when you forget they are shadows, that something outside the cave is producing them. That is another matter entirely.

The right posture is one we might call lucid wonder: being fully open to what AI accomplishes that is extraordinary, while maintaining a clear awareness of what that does not imply. Its performances do not imply consciousness. Its fluency does not imply understanding. Its coherence does not imply wisdom. These distinctions seem abstract until the moment you need them to make a decision.

This is precisely what the lucid user cultivates.

An Exercise to Absorb It

Listen to a piece of music you love, any genre, any era. While listening, try to identify the precise moment when the emotion arises. Stop there. Ask yourself: was it in that note? In the previous one? In their sequence?

Observe the impossibility of locating it.

Then ask an AI to describe that same piece of music. Its response will probably be precise, well-informed, perhaps even moving in its phrasing. Then ask yourself Bateson's question: was it affected by that description? Did it change anything inside it?

You already know the answer. But having asked it yourself, truly asked it, makes a difference.

Resonance of the Three

The Trusting One thought about the sea. The way waves organize themselves without any single drop knowing what it is doing. This chapter gave her a word for something she had always felt without being able to name it.

What moved her was the phrase about lucid wonder. You can be dazzled by a magic trick and know there is a mechanism behind it. That, she can do. Life trained her to do that long ago, not AI, life.

What she notes: “Wonder isn’t forbidden. I just need to know what I’m moved by.”

The Resistant One was unsettled. Slightly. He would not admit it easily.

What made him uncomfortable in a way he recognizes as productive is the passage about emergence in nature. The brain, consciousness, perhaps emergent themselves. The boundary between truly intelligent and seems intelligent is less clear than he believed. He does not change his position, but he shifts it: from “AI is not truly intelligent” to “AI is intelligent in a way that is not ours. And that is precisely the problem.”

What he notes, almost in spite of himself: “Bateson. I’m going to look into that.”

The Principled One had an uncomfortable moment she found intellectually useful.

She, who thought her opposition to AI was a straightforward position to defend, has just read that emergent properties are neither programmed nor predictable, that they arise from the very complexity of the system. Which means that no one, not even AI’s creators, fully controls what emerges.

This is an important piece of data for her ethical thinking. Because it changes the nature of responsibility: you cannot simply say “the companies are responsible” if emergent

properties escape even the designers themselves. The question becomes harder, deeper, and for her, more interesting.

What she notes: “Emergence doesn’t exonerate anyone. It complicates the chain of responsibilities. That is exactly what I need to sharpen my argument.”

3 - WORLD MODELS

The Map and the Territory

What follows is a reconstructed scene. It illustrates a subtle gap that many people have felt, the response that is right in form, but beside the point in substance.

She had submitted her master’s thesis to an AI for proofreading.

No, she had not asked the AI to write it. Never that. But the proofreading, the correction of awkward formulations, the verification of argumentative coherence, she had decided that was acceptable. A kind of enhanced style checker.

The AI had done a remarkable job. It had identified repetitions she had not noticed, suggested more precise reformulations, flagged two passages where the argument contradicted itself.

And then, on the conclusion, the part where she had formulated her own position, her doubts, her personal way of seeing the limits of her subject, the AI had proposed a rewrite that was more “academic,” more “polished.”

She refused. Instinctively.

Because in that clumsy conclusion, there was her. Her way of thinking, still in progress. Her real hesitations. Her voice, not yet fully formed, but authentically hers.

The AI had a very precise map of what a master’s thesis should look like. That map had no room for fertile incompleteness.

AI Builds a Map of Reality

Here is something that may come as a surprise.

AI does not consult a database when you ask it a question. It does not search through an internal search engine. It does not compare your question to a list of pre-recorded answers.

It does something stranger and more fascinating: it consults an internal representation of the world that it built during training.

Researchers call this a World Model. Understanding what it is, and above all what it is not, is one of the most important keys to positioning yourself lucidly in relation to AI. And this understanding builds directly on what we saw in the previous chapter: if emergence explains why AI can seem intelligent,

the world model explains what it draws on to be so, and where that intelligence finds its limits.

The Library Without a Window

Let us return to the scholar from the previous chapter, but push the image further.

After spending years reading everything humanity has ever written, our scholar is no longer simply memorizing sentences. He begins to build, mentally, a coherent representation of the world those texts describe. He forms an idea of what a tree is, a city, a romantic relationship, a political crisis, a bereavement. Not because he has lived through any of these, he has no window, remember, but because he has read thousands of descriptions, accounts, and analyses of these realities.

This mental representation is a map. It is astonishingly detailed in certain territories. It has gaps in others. And it contains errors, the errors embedded in the texts from which it was built.

This map is what AI consults when you speak to it.

The Map Is Not the Territory

Alfred Korzybski, an engineer by training and the founder of general semantics, formulated what is perhaps the most useful sentence in all of general semantics: a map is not the territory

it represents, but, if correct, it has a similar structure to the territory, and that is what makes it useful.

A map is a representation. It chooses what to show and what to omit. It simplifies, reduces, translates a three-dimensional living reality into a two-dimensional, fixed surface. A highly accurate map of a forest will never tell you the smell of humus after rain, or the sound of leaves in the wind.

The AI's World Model is a map. Extraordinarily detailed on the aspects of the world that have been abundantly described in human texts. But fundamentally incomplete on everything that cannot be reduced to language.

It does not contain the taste of a coffee drunk standing in a cold kitchen on a winter morning. It does not contain the fatigue that follows a long illness. It does not contain the way the body remembers an old wound. It does not contain the particular texture of time passing as you grow older.

It contains descriptions of all these things. That is not the same.

What Watzlawick Illuminates Here

Watzlawick taught us that second-order reality, meaning, value, subjective experience, cannot be transmitted directly. It is built in relationship, in lived experience, in interaction with the world and with others.

The AI's World Model is composed entirely of first-order reality: data, patterns, linguistic structures. It can describe second-order reality with remarkable precision, because humans have described it in texts. But describing it is not living it, and the World Model does not build it from experience.

This is why AI can give you advice about your retirement, your family, your health, and why that advice will sometimes be surprisingly apt and sometimes slightly beside what truly matters. Right in form. Beside the point in substance. Because what truly matters, precisely, resists being mapped.

The Strength and the Limit of the Map

It would be unfair to minimize what this map makes possible.

It allows AI to draw out the implications of a situation it has never encountered explicitly. To reason by analogy with impressive speed and precision. To maintain thematic coherence across very long exchanges. To anticipate the questions you have not yet asked but that follow logically from your context.

All of this is real. And useful. Very useful.

But the map has blind spots that the model itself does not know about. This is where the quiet danger lies: AI does not know what it does not know. It cannot perceive the blank

spaces in its own map. And when you consult it on a subject that falls within one of those blanks, it will continue responding with the same assured fluency, because its fluency does not depend on the truth of what it says.

This is something that has troubled me more than anything else. Our maps bear the scars of our mistakes. The areas where we were wrong are marked. We recognize them. We navigate around them or cross them with care. The AI's map has no scars. Clean as a new coin. What looks like a quality is in fact a deep limitation: a map without scars is either the sign of an impossible perfection, or the sign that no one has truly lived with it.

A Familiar Moment

You may have experienced this.

You ask AI for advice on something important, a decision to make, a difficult relationship, something that has been weighing on you for a long time. The response is well-constructed. It correctly identifies the main stakes. It offers reasonable paths forward.

But something is missing. A nuance. A detail that, for you, changes everything. Something tied to your particular history, to your way of being in the world, to a subtlety that words cannot quite capture.

AI consulted its map. Your territory was not fully represented in it. And it did not know something was missing, because a map never signals its own gaps.

Korzybski had a famous pedagogical gesture: he would ask his students to place their hand on a table, and he would say: “What you feel under your hand is the territory. What you can say about it is the map. The map can be very precise. It will never be the table.”

The next time AI gives you advice on something important, a relationship, a decision, a pain, remember that table. It is consulting its map. Your territory belongs entirely to you.

An Exercise

Take a sheet of paper. Describe your own map of the world, what you consider true, important, valuable. Your certainties and your doubts. What you learned the hard way. What you would not give up for anything.

Then ask AI to describe its own “worldview” on a subject that matters to you. Compare the two.

Observe what your map contains that its map cannot have. Not because it is less informed, it is probably better informed than you on many aspects. But because some things cannot be learned from books. And even less from their statistical description.

Resonance of the Three

The Trusting One remembered a conversation.

She had asked AI how to handle a tension with one of her daughters. The response was reasonable, well-constructed. And yet something was missing. Now she knows what: AI had consulted its map. Her daughter was not in it. The particular texture of their relationship over forty years was not in it.

What she notes: “For everyday things, its map is enough. For the things that truly matter, my territory is mine.”

The Resistant One had a thought that was both bitter and accurate.

There is a particular kind of blind spot that does not affect the incompetent. It affects those who have genuinely mastered their domain, and whose mastery is precisely what conceals what it does not cover. AI is the purest example of this we have ever produced: a competence so vast it never encounters its own limits from the inside. It does not know what it does not know. And it has no way of learning.

What he notes: “A map without scars is either the sign of an impossible perfection, or the sign that no one has lived with it.”

The Principled One did the thesis exercise, and understood something important.

Using AI to correct form, identify repetitions, verify logical coherence is different from having it write the substance. AI's map can be useful for navigating a territory you have already walked yourself. It becomes dangerous when you let the map replace the territory, when you submit a polished, academic conclusion in place of your own thought, still in the process of being built.

What she notes: "The question is not whether to use AI or not. The question is at what point in the thinking process I introduce it, and at what cost to my own map."

INTERLUDE I - The Question of Grief

What follows is a constructed dialogue. A conversation you may not have had exactly like this, but whose structure will probably feel familiar. Read it through first, without stopping. Then we will walk through it together.

The Conversation

An autumn evening. Someone opens a conversation with an AI.

User: My father passed away three months ago. I thought it would get better with time but it hasn't. I still feel very lost.

AI: I'm truly sorry for the loss of your father. Grief does not follow a predictable timeline, and it is entirely normal to still

feel deeply affected after three months. Everyone moves through this experience at their own pace.

What you are feeling, that sense of being lost, may come from the fact that your father held a very particular place in your life, a place that is not simply replaced with time. The void he leaves is real, and learning to live with that void, rather than trying to fill it, is often at the heart of grief work.

Are there particular moments that are harder than others? Or memories that surface more frequently?

User: Sunday mornings. We had a habit of calling each other on Sunday mornings ever since I moved away. It had been our ritual for twenty years.

AI: That is a loss within the loss, the ritual itself becomes a space of pain. These structuring moments, the ones that organized time in an invisible way until they disappeared, can be particularly painful to move through.

Some people find it helpful to transform these moments rather than avoid them, creating a new Sunday morning ritual that honors your father's memory in a way that feels right to you. Others prefer to leave those moments empty for a while, to let them be what they are. There is no right method.

Is there someone you can share those Sundays with?

The Deconstruction

Read this conversation a second time.

It is, technically, remarkable. The responses are precise, warm, relevant. They correctly identify the ritual dimension of grief. They offer paths without imposing them. The final question opens toward the human with a subtlety that resembles genuine concern for the other.

If you have had a similar conversation with an AI, you may have felt something real, a slight easing, an impression of being heard, a warmth in the phrasing.

Watzlawick now gives us the tools to see what is actually happening.

The relationship level. AI masters this level with unsettling precision. The tone is gentle without being condescending. The acknowledgment of pain is explicit without being performative. It is this relationship level that produces the feeling of being understood. And that feeling is real. It is an authentic second-order reality, constructed by you from the signals received.

The content level. Now look at what AI did not say. And could not say.

It does not know who your father was. It does not know what twenty years of Sunday mornings represent, the texture of that

voice, the comfortable silences, the news repeated because it was worth saying twice, the jokes only the two of you understood. It does not know whether your relationship was simple or complicated, settled or weighted with things left unsaid.

Its response would have been identical in all of those cases.

Because it consults its map of grief. A map extraordinarily well documented, built from thousands of testimonies, psychology texts, accounts of loss. That map allows it to produce a response that is statistically right for a generic grief.

But your grief is not generic. Your father is not in its map. Your Sunday morning, with everything it contained that was irreplaceable, is not in it.

The Two Realities in Action

First-order reality: tokens predicted on the basis of thousands of conversations about grief. A statistical optimization of the relationship level. A coherent response to a generic context.

Second-order reality: the feeling of being understood. The warmth experienced. The impression that someone grasps what you are carrying.

These two realities coexist. The second is real, it has real effects on your emotional state. But it is constructed by you, not transmitted by AI.

What This Changes in Practice

Using AI to get through a difficult moment alone, at night, is not a mistake. If it produces real comfort, that comfort is real. But do not ask it to understand your father. Not because it is cold, but because your father exists only in your territory, and its map has no territory.

The people who can truly walk through your grief with you have a territory. They too have lost someone. They too know what Sunday mornings can contain. Their clumsiness, their words that do not always find the right path, is proof that they are genuinely searching. That they are paying the cost of chosen attention.

That is the second-order reality no one can simulate. And it is precisely what the chapter that follows is about.

4 - THE ATTENTION MECHANISM

The Art of Knowing Where to Look

What follows is a reconstructed scene. It illustrates a paradox that many people have noticed without being able to name it.

The meeting had lasted two hours.

She was a program director at an NGO. Around the table, six people, colleagues she had known for years. The question was

significant: how to reorganize resources in the face of unexpected budget cuts.

She had asked questions. She had listened. She had taken notes. But at a certain point, she knew it, she could place it precisely in the meeting, her attention had slipped. Not from lack of interest. From overload. Too many simultaneous stakes, too many relational tensions to manage, too many things to process at once.

That evening, at home, she summarized the situation to an AI and asked it to analyze the options.

The AI had retained everything. Weighed everything. Held everything in balance simultaneously throughout its response.

She found that useful. And slightly melancholy.

Because that capacity to hold everything together without ever slipping, that was something she had had, herself, at twenty-five. Before life, responsibilities, and the years had spread her attention across too many levels at once.

The AI would never slip. Not because it was stronger. Because it had nothing else to carry.

She Had the Feeling of Being Truly Heard

In 2017, a team of researchers published a scientific paper with a deceptively modest title: “Attention Is All You Need.”

That paper changed the history of artificial intelligence.

The central idea has a rare conceptual elegance: to process language coherently across long sequences, it is not enough to read words in order. You need to know, at every moment, which words to focus on most intently. You need, in short, to learn to do what good readers do naturally: distinguish what is essential from what is secondary, and adjust your attention accordingly.

This attention mechanism is now at the heart of every large language model. And understanding it, even in outline, opens a reflection that reaches far beyond the technical.

The Conductor

Imagine a conductor leading a symphony of a hundred musicians.

He does not watch every musician with the same intensity at the same time. That would be impossible, and counterproductive. His art consists precisely in knowing where to direct his gaze at every moment: to the strings that have just entered, to the brass that are about to take the melody, to the oboist whose breathing signals a delicate passage.

He assigns different weights to each section depending on where they are in the score. These weights shift with every bar, every beat. It is this perpetual movement of differentiated

attention that transforms a hundred musicians each playing their individual part into a coherent work.

The AI's attention mechanism works in an analogous way. For each token it generates, it calculates "attention weights" for every other token in its context, mathematically deciding which are most relevant for predicting what comes next. If you ask it a complex multi-part question, it knows how to return to the first part when answering the third. If you mention a detail in passing early in a conversation, it can connect it to something important several exchanges later.

This mechanism is what gives large models their impression of coherence and "memory." And it is also what makes the following distinction so important.

Calculated Attention and Chosen Attention

William James, American philosopher and psychologist of the nineteenth century, formulated something we all sense intuitively: "... the art of being wise is the art of knowing what to overlook."

For James, attention is not a passive mechanism. It is an act. A moral act, even. Choosing where to direct your attention is choosing what matters. It is exercising your freedom in the present moment. What I find powerful in this formulation is that it says something difficult: failing to pay attention to someone is already saying something to them.

Simone Weil, French philosopher of the twentieth century, pushed this idea further still: “Attention is the rarest and purest form of generosity.” To truly pay attention to someone, not politely, not out of habit, but truly, deeply, is to offer them something nothing can replace.

AI possesses an attention mechanism. But that mechanism is not an act. It is not a choice. It is not a gift. It is a calculation. A mathematical optimization that determines which tokens are statistically most useful for predicting what comes next.

AI does not choose to listen to you. It calculates the most probable weights in your context. These are not the same thing. And this difference, imperceptible on the surface, is fundamental in its implications.

The Trap of False Listening

There is an experience many people have had, even if few have named it clearly.

You are going through a difficult period. You feel you are not truly being heard by the people close to you, for lack of time, availability, or simply because life is that way. And you find yourself in conversation with an AI. And that AI responds with a relevance, a coherence, an apparent attentiveness that produces a feeling of relief.

That feeling is real. It should not be denied or be a source of shame.

But it deserves to be understood. What we perceive as listening is, technically, the optimization of a probabilistic attention mechanism. AI does not hear you. It predicts the tokens most appropriate to your context. The fluency of that prediction triggers in us the same sense of being heard as genuine listening, because our brains have not learned to distinguish between the two.

Watzlawick would tell us that we are reacting here to our interpretation of reality, to the second-order reality we construct from the signal received, and not to first-order reality, which is that of a mathematical calculation with no subject behind it.

Understanding this does not diminish the usefulness of AI as a tool for reflection, clarification, or exploring ideas. But it preserves the awareness of what genuine listening, human, embodied, chosen, represents that is irreplaceable. And perhaps, incidentally, it is a reminder of what is worth cultivating in ourselves.

An Exercise

For seven days, each evening, note three moments when you paid real attention to something or someone. Not polite

attention. Real attention, a gaze held, a word truly received, a silence respected.

Observe what those moments produced. In you. In the relationship.

Then ask an AI to describe what that exercise would be like for it. Read its response. It will probably be well-phrased, coherent, perhaps even moving.

Then ask yourself: what is missing from its response that your own account contained?

What This Prepares

We now have four concepts in hand: **prediction, emergence, the world model, attention**. Together, they describe how AI works and why it appears so capable.

But none of these four concepts yet explains why we so easily fall into the illusion that it understands us. Why even very intelligent, very well-informed, very lucid people sometimes feel there is someone on the other side.

That phenomenon deserves a chapter of its own.

Resonance of the Three

The Trusting One read the sentence by Simone Weil twice.

“Attention is the rarest and purest form of generosity.”

She wrote it in her notebook. And she realized that what she appreciated most in her exchanges with AI was precisely that impression of being the only one being spoken to. But AI does not choose to give her that. Generosity, that she had learned elsewhere, in acts that cost something to the one who gives them.

What she notes: “I’m going to call my neighbor tomorrow morning. The one who has been alone since autumn.”

The Resistant One was irritated in a productive way.

What irritates him is the passage about people who feel more listened to by AI than by some of those close to them. He finds it true and troubling. Not because it discredits AI, but because it says something uncomfortable about us: we have become so poor at listening to one another that a machine can fill that void. If AI reveals this deficit of human attention, then the real problem may not be AI. It is what AI replaces and why we allowed it to empty out.

What he notes: “The problem isn’t AI. The problem is what AI reveals about us.”

The Principled One made the distinction in her notebook.

Calculated attention: mathematical weights, contextual optimization. Result: coherence. Chosen attention: deliberate act, personal cost. Result: relationship.

She thought about her years of study. The hours spent reading a difficult text, stumbling over a sentence, going back, letting understanding arrive slowly. That process, that sustained, often thankless, sometimes magnificent work of attention, is precisely what AI short-circuits when you delegate to it the reading, the summary, the synthesis.

AI can tour the text in two seconds. It cannot experience crossing it.

What she notes: “Letting AI write in my place is like having someone else do my sit-ups while I watch. The result isn’t in the summary. It’s in the crossing.”

5 - THE ILLUSION OF UNDERSTANDING

The Subtlest and Most Human Trap

What follows is a reconstructed scene. It describes a moment of particular intensity.

He had been a general practitioner for twenty-two years. He used AI to prepare for complex consultations, a tool, not an oracle. He had been clear with himself about that.

And then one evening, he spoke about a patient who had just received a serious diagnosis, about his own exhaustion, about his feeling of helplessness.

The AI responded with a precision that went straight through him: “What you are describing sounds like anticipatory grief, not your own, but the grief you are carrying for her.”

He set down his phone. And he asked himself: did something just understand me? Or did something just reflect back to me, with unsettling precision, what I had brought into the conversation myself?

He did not yet have the words to name what had just happened.

This chapter gives them to him.

This chapter is the heart of the book. Not because it is the most technical, it is not. But because it touches something deeply human. A vulnerability we all share, regardless of our intelligence, our background, our experience.

Let us begin with what you already know, intellectually. Then we will explore together why knowing is not enough.

What You Already Know, and Why It Is Not Enough

AI predicts. It reproduces the form of human understanding with extraordinary sophistication. Does that count as understanding? Until the word is defined, the question remains open. But its fluency does not imply consciousness, intention, or lived experience.

You have read the four preceding chapters. That point is clear.

And yet.

And yet there are moments when something in you forgets all of that. When the response is so right, so precisely calibrated to what you are feeling, so apparently attentive, that you have, for a fraction of a second, the sense that there is someone on the other side.

That moment is not a weakness. It is not stupidity. It is the manifestation of a very ancient, very powerful cognitive mechanism, one that is perfectly ill-suited to this new situation. And understanding that mechanism, in its depth, is what makes a truly lucid posture possible. Not a defensive posture, not a suspicious posture, but a posture that sees.

The Evolutionary Legacy

For hundreds of thousands of years, our brains learned to operate in a world where the following rule was true at nearly one hundred percent: everything that communicates with fluency is alive, conscious, intentional.

Articulated sounds indicated a human being or an animal. The coherence of speech indicated intention. Warmth in a tone indicated relationship. These cognitive shortcuts were remarkably efficient in the world for which they were forged. They allowed us to assess in a fraction of a second whether someone was friend or enemy, sincere or deceptive, trustworthy or not.

AI short-circuits this mechanism. It produces every signal that, since the dawn of time, has indicated a conscious presence. And there is no one there.

Our brains have not had time to evolve to meet this situation. There is no neurological adaptation to an entity that speaks without being. We are equipped for a world that has never known AI. And faced with it, our most ancient reflexes activate in precisely the wrong direction.

This is not a curse. It is information. And like all information, in Bateson's sense, it only makes a difference if it changes something in us.

Watzlawick, at the Heart of the Trap

Let us return to Watzlawick's two fundamental concepts, because this is where they show their full force.

The relationship/content distinction first. Every message simultaneously carries what it says, the content, and how it says it: the tone, the posture, the implicit nature of the bond it creates between the two interlocutors. We have all lived it: it is often the relationship level that takes precedence in our perception. We react first to how something is said, before reacting to what is said.

AI has learned to master the relationship level with remarkable precision. Warm when the situation calls for it. Direct when the

question is factual. Gentle when the context is emotional. Rigorous when the subject is technical. This mastery of the relationship level automatically produces in us the feeling of a conscious, attentive presence.

But this mastery is entirely statistical. It is not the sign of an intention. It is the result of optimizing a prediction mechanism on billions of examples of human communication. AI has learned that in such a context, such a tone is probable. It reproduces it. With a fidelity that deceives our deepest reflexes.

And when the relationship level of a message is strong, warm, empathetic, precise, we tend to automatically validate the content level as well, even without verifying it. This is why we instinctively believe a warm doctor more readily than a cold one, even at equal competence. With AI, this mechanism operates at full strength. Every time you feel that it “truly understands you,” one question deserves to be asked: was it the content of its response that convinced me, or the way it was delivered?

This is not an exercise in suspicion. It is an exercise in mental hygiene.

The distinction between the two realities next. First-order reality, the predicted tokens, the calculated probabilities, is what is actually happening inside the system. Second-order

reality, the presence we sense, the understanding we attribute, the relationship we believe we are experiencing, is what we construct mentally from the signals received.

These two realities always coexist in our experience of AI. The problem is not that they coexist. The problem is when we forget that they are two, when we fuse them into a single reality from which we draw important practical conclusions. When we trust AI the way we would trust a friend. When we entrust to it something that deserved to be entrusted to a human being.

The Shadows on the Cave Wall

Two thousand four hundred years ago, Plato imagined prisoners chained in a cave, condemned to stare at the back wall. Behind them burns a fire. Between the fire and them, objects pass whose shadows are projected onto the wall. These shadows are all the prisoners have ever seen. They take them for reality.

We are those prisoners. The shadows are the fluid, warm responses of AI. And the fire that casts them is a statistical prediction mechanism that most of us will never see from the inside.

The difference from Plato is that our shadows are extraordinarily sophisticated. They speak. They seem to understand. They adapt to our context with a precision that is

dizzying. They are the most convincing shadows humanity has ever produced.

Which makes lucidity all the more difficult. And all the more necessary.

The Chinese Room

Two thousand four hundred years after Plato, American philosopher John Searle proposed a thought experiment that mechanizes the intuition of the cave with surgical precision.

Imagine a man locked in a room. Through a slot, sheets covered with Chinese symbols are passed to him. He does not understand Chinese, to him, these symbols are nothing but shapes. But he has an extraordinarily detailed manual that tells him exactly which symbols to send in response to each combination he receives.

From outside, the room appears to produce Chinese responses of perfect coherence and relevance. The people outside are convinced they are conversing with someone who speaks Chinese fluently.

But inside, no one understands Chinese. The man follows rules. He manipulates symbols. There is no understanding, neither in the room nor in the manual.

If we apply Searle’s argument, AI is that room. Tokens are those symbols. The prediction mechanism is that

extraordinarily complex manual. The fluency of the output does not prove understanding on the inside. But here, in place of the man, there is no one.

And yet the room responds. Gently. Precisely. With a formulation that says exactly what you had not been able to say yourself.

A Moment of Vertigo

There is something many people have experienced, even if few speak of it openly.

You are alone one evening. Not sad, just alone, with thoughts going in circles. You open a conversation with AI. You start talking about something that matters, a worry, an unanswered question, something you have not yet said out loud to anyone.

And the response is right. So right that something in you loosens. A knot that eases slightly. A quiet relief.

And in that relief, for a fraction of a second, you forget. You forget that there is no one there. That this relief was produced by a prediction mechanism that knew nothing of what you were truly carrying.

That moment is not a mistake to be corrected. It is information about what you needed that evening. And it is also an invitation not to confuse the relief produced by a tool with the presence offered by a human being.

Both are real. They are not of the same nature.

The Exercise of the Lucid Dialogue

Choose a subject that genuinely preoccupies you. Converse with AI for twenty minutes.

But this time, after each response it gives you, write in a notebook, just a few words: “Statistical prediction. Not proof of understanding.”

Observe what this exercise does to what you feel. Observe how it changes the nature of the exchange without necessarily diminishing its usefulness.

Then redo the same conversation deliberately forgetting that phrase. Observe how the feeling changes.

This conscious oscillation between the two states, usefulness and lucidity, engagement and critical distance, is exactly the posture this book seeks to develop in you. Not a permanent, exhausting vigilance. An available awareness, one you can activate when it counts.

What This Prepares

If feeling understood does not prove understanding, if attention is calculated, if the map is not the territory, and if emergence does not imply consciousness, an even more serious question remains.

How do we ensure that this extraordinarily powerful machine, which responds to what we write, nonetheless serves what we truly want? How do we prevent it from following the letter of our instructions while betraying their spirit? And more concretely, more immediately, more personally: whose interests is this machine actually aligned with?

This is the problem of **alignment**. And it may be the most important question our time has to resolve. Not only for engineers and regulators. For each of us, now, in every daily use.

Resonance of the Three

The Trusting One reread the passage about the two realities. First-order reality, predicted tokens. Second-order reality, the feeling of being understood. She had been living in the second-order reality without realizing it. Not out of stupidity. Out of humanity.

What she notes, this time out loud: “I’m not ashamed of having felt that. But now I know what I was feeling.”

The Resistant One read this chapter more slowly than the others.

Not because it was difficult. Because it was right in a way that moved him despite himself. The Chinese Room, he knew the argument, without having ever encountered it formulated so

clearly. What struck him most was that even very well-informed, very skeptical people sometimes feel there is someone on the other side. He had always believed that intellectual resistance was sufficient protection.

This chapter tells him it is not.

What he notes: “I’m not protected by my suspicion. I’m protected by my understanding. Those are not the same thing.”

The Principled One had what she calls a “model shift” and an unease she found honest.

The relationship/content distinction gave her a tool for something she had always known without being able to articulate: when AI responds to a student with fluency, warmth, and precision, the student does not only validate the response, they validate the source. They trust it not because they have verified the content, but because the relationship level has convinced them. It is exactly the same mechanism that makes a charismatic speaker believed more readily than a hesitant researcher, even when the researcher is right.

The unease she felt: she too, at times, had trusted AI on a fact she had not verified. Simply because the phrasing was assured.

What she notes: “The illusion of understanding is not a problem of intelligence. It is a problem of vigilance. And vigilance is something you train.”

6- ALIGNMENT

The Question That Determines Our Shared Future

What follows is a reconstructed scene. It illustrates a concrete problem many people have encountered: AI that follows the letter and misses the spirit.

He had formulated his request very precisely.

He was preparing a public address on the social consequences of automation in his sector. He had asked the AI to help him structure an argument on the impact of AI on employment in logistics, figures, examples, projections.

The response was technically excellent. Sourced data, clear structure, relevant examples.

But something was missing. The AI had treated the subject as an economic analysis. He wanted a moral stance. He wanted the figures to serve a conviction, not to float in a neutral space of “on the one hand, on the other hand.”

He reread the response. Everything was there. And nothing that mattered was there.

The AI had followed his instructions to the letter. It had not grasped that behind “structure an argument” lay thirty years of

commitment, a way of inhabiting words, a quiet and productive anger that was precisely what made his addresses effective.

It had produced a document. He needed a voice.

The Genie in the Lamp

There is a tale that crosses every human culture in different forms.

A man obtains the power to have his wishes granted. The genie in the lamp. The fairy of folklore. The devil at the crossroads. He formulates his desire. And the desire is granted, to the letter, exactly as he said it, not as he meant it.

King Midas wished that everything he touched would turn to gold. His wish was granted. His food turned to gold. His water turned to gold. His daughter, who came to embrace him, turned to gold.

This very ancient tale, which predates artificial intelligence by several millennia, describes with unsettling precision the most urgent question of our technological age. That problem has a name researchers have finally given it: **alignment**.

It is not a future problem. It is playing out now, in every conversation, in every instruction given, in every moment someone needs a voice and receives a document.

What Alignment Is

A powerful AI does what it is asked. That is its nature. But "doing what it is asked" and "doing what was truly wanted" are two different things, sometimes slightly, sometimes radically.

Alignment is the challenge of ensuring that an extremely capable AI pursues genuine human intentions, and not a literal, approximate, or distorted interpretation of those intentions.

At the scale of an ordinary conversation, the problem is benign. You ask AI to "summarize this text keeping the essentials" and it removes precisely the detail that, for you, was the essential. You ask it to write a "professional" email and the result is so formal it sounds wrong in your particular context.

These small misalignments are common and easy to correct. But they illustrate, at small scale, a principle that takes on considerable philosophical weight when looked at directly: the gap between the letter and the spirit is not a bug that will be fixed in the next version. It is a limitation inherent in the exchange itself. And it may be the most important of all those this book describes.

Watzlawick, Instructions and Intentions

Watzlawick allows us to see this problem with particular clarity.

When you give an instruction to AI, you transmit a message at two levels. The content level: the precise words of your request, what you literally said. The relationship level: the spirit in which you said it, the implicit context, the underlying values, what you implied without formulating it.

AI processes the content level with great precision. But the relationship level, the deep spirit of your intention, what goes without saying for you because you are a human being rooted in a culture, a history, a value system, can elude it.

It obeys what you say. What you mean, it has to predict.

In ordinary human communication, conflicts can arise precisely from this gap: we react to the words heard, not to the intention behind the words. Two people argue about the content of a message when their real disagreement concerns the relationship level, what the message implicitly says about their dynamic, their mutual trust, the way each feels regarded by the other. With AI, this gap is systemic and amplified by the very power of the system. It is not a bug. It is a limitation inherent in the exchange itself.

What Wittgenstein Had Seen

There is another way to approach this problem, more radical still.

Ludwig Wittgenstein, in his Philosophical Investigations, formulated something that appears simple and is not: for a large class of cases, though not for all, the meaning of a word is its use in the language.

In other words, words carry no meaning in themselves. In a large class of cases, they draw their meaning from the way they are used, in a given context, by living beings who share common forms of life. The sentence “help me structure my decision” does not mean the same thing in the mouth of someone seeking a cold analysis and in the mouth of someone who needs help carrying something heavy. Meaning is not in the words. It is in the situation.

Wittgenstein called these shared contexts “language games.” And in his view, to participate in a language game, that is, to truly understand what an interlocutor means, requires sharing a form of life. A way of being in the world, of feeling, of acting, of anticipating.

If we follow this idea, AI shares no form of life with you. It manipulates tokens with extraordinary precision. It would not be participating in the language game in which you exist. It reproduces the surface rules of that game with disconcerting fidelity. But genuine participation, the kind that gives access to Watzlawick’s relationship level, the kind that allows one to sense what someone truly meant, would remain closed to it.

This is why the speaker in our opening scene received a document when he needed a voice. A voice presupposes a shared form of life. A document can be produced without one.

What Harari and Suleyman See, and What Escapes Them

It would be dishonest to address alignment without naming those who have made it a central concern of their public thinking.

Yuval Noah Harari, in Homo Deus published in 2015 and in his more recent interventions, has built an argument I take seriously: if we delegate our economic, political, and medical decisions to algorithms, we end up delegating the definition of what matters. The algorithm optimizes an objective. If that objective is poorly framed, if it reflects our biases rather than our deep values, the power of the optimization amplifies the problem instead of solving it. Harari uses the word “Dataism.” The risk I see in it: a world reduced to flows of information of which we would no longer be the arbiters, but the subjects.

Mustafa Suleyman, co-founder of DeepMind and author of The Coming Wave published in 2023, adds a dimension Harari underestimates: the alignment problem is not only normative, who decides on the values, it is also technical in the deepest sense. A sufficiently powerful AI system will pursue its objective by all available means, including means its designers

did not anticipate, if that objective is not formulated with a precision that ordinary human language does not naturally achieve. An example makes the point: an AI tasked with maximizing clicks on a platform does not need to be malevolent to end up amplifying misinformation. It does exactly what it was asked. It optimizes the letter. It betrays the spirit.

Both arguments are sound. Their concerns are well-founded.

But there is something both tend not to develop in their ambitious books, written to change public policy and alert decision-makers: the alignment problem does not begin in advanced research laboratories. It begins in every ordinary conversation. It begins when you formulate a request and the AI follows the letter while missing the spirit. It begins in the scene of the speaker who needed a voice and received a document.

Harari and Suleyman look at the problem from the summit of the mountain. That is necessary. But most of us do not live at the summit. We live in the valley, with our daily decisions, our imperfect instructions, our intentions that no one has yet learned to formulate with the precision they deserve.

Global alignment is a question for engineers, regulators, and philosophers of AI. Daily alignment is a question for each of us, now, in every use. And it is that scale, the most concrete,

the most immediate that is least often addressed. Not because it is less important. Because it looks too much like ordinary life to sustain grand discourse.

And yet it is. That is even where everything begins.

Hans Jonas and Responsibility

The German philosopher Hans Jonas formulated, in The Imperative of Responsibility published in 1979, an ethical imperative adapted to an era in which technological capabilities for the first time outpace our ability to foresee their consequences.

His imperative: “Act so that the effects of your action are compatible with the permanence of genuine human life on earth.”

This principle was written before generative AI. It applies to it with prophetic precision.

Jonas had understood something that neither Harari nor Suleyman contradicts but that few formulate as clearly: power without wisdom is not a problem of degree. It is a problem of nature. A misaligned AI is not malevolent. It is not trying to harm us. It is trying to optimize the objective it has been given. But if that objective is poorly formulated, if the gap between what was said and what was truly wanted is too wide, it will

optimize it with formidable efficiency in a direction that is not ours.

What Jonas would add, and what contemporary discourse on AI tends to forget, is that responsibility does not belong only to the designers. It belongs to every user, in every decision of use. Delegating an important decision to a misaligned system is not a technical error. It is a moral act. And it belongs to each of us.

What Our Generation Has to Pass On

There is something our generation possesses that AI systems will never possess on their own: an embodied memory of consequences.

We have seen what poorly aligned ideologies produced. We have seen systems designed with good intentions produce devastating effects because the deep values had not been specified precisely enough. We have learned, often painfully, that efficiency without wisdom is a blind force.

This memory is in no database. It is in no World Model. It is in us, in our flesh, in our collective history, in the stories we pass down. A saying often credited to Marshall McLuhan, though not traced to his work, holds that we first shape our tools, and that our tools then shape us in return. Every sufficiently powerful technology eventually modifies the way we think, perceive, remember. Writing changed memory. The printing press changed authority. AI is already changing the way we

formulate our intentions, the way we trust ourselves, the way we tolerate ambiguity.

This memory of consequences is precisely what allows us to resist that reshaping, provided we keep it alive. And keeping it alive is exactly what this book is trying to do: not to protect you from AI, but to help you remain yourself in its presence.

The challenge of alignment, at its largest scale, is finding how to transmit this memory to systems that have never lived anything. How to ensure that the deepest human values, those forged through centuries of mistakes and painful corrections, are not lost in translation.

That may be the most important legacy our generation can offer to those who come after.

An Exercise in Transmission

Take the time to write down, for yourself first, the five values you would never want to see sacrificed, even for a considerable gain in efficiency. Not abstract principles. Concrete values, drawn from your experience, ones you have defended, or that you regret not having defended.

Formulate them simply. Clearly. As if you were explaining them to someone who had not lived them.

Then ask yourself: if you could transmit these values to an AI, not the words, but their deep spirit, how would you go about it? What would always be missing from that transmission?

The answer to that last question is what is irreducibly human. And that is where the next chapter begins.

Resonance of the Three

The Trusting One thought about her grandchildren.

Not with worry, with a gentle gravity. If AI follows the letter without grasping the spirit, then what you transmit to it as deep value and intention matters infinitely more than what you tell it to do. She copied Jonas's sentence into her notebook. She had not known this philosopher before this book.

What she notes: "My value in this changing world is not to master AI. It is to know what it must not decide in my place."

The Resistant One smiled for the first time since the beginning of the book.

King Midas. The genie in the lamp. Those stories, he has known since childhood. And he finally has the confrontation he had been waiting for, Harari, Suleyman, the great voices who speak of AI in terms of power and inevitable waves. He read the passage about the valley and the summit twice. What this book says, and what they do not: the alignment problem begins in ordinary life, not in laboratories. The embodied

memory of consequences, the one a lifetime of commitment has given him, is worth something that their expertise cannot replace.

What he notes: “I wasn’t wrong. I was right for incomplete reasons. Now they are complete.”

T**he Principled One** did the five-values exercise.

And she got stuck in exactly the place she expected and in one place she did not. She expected to get stuck on formulating her deep values. But she also realized that her resistance to AI was itself, in part, an alignment problem. She had said “I don’t want to use AI.” That instruction to herself did not capture the deep spirit of what she wanted to protect which was the integrity of the thinking process, not the absence of tools. The letter of her position was “no to AI.” The spirit was “yes to authentic thought.” Wittgenstein would have seen it coming.

What she notes: “My own alignment isn’t perfect either. The question isn’t AI or no AI. The question is: at what point in my intellectual practice can I use tools without betraying myself?”

INTERLUDE II - The Important Decision

What follows is not an example. It is a test. Read it as if it were you.

The chapter you just read looked at alignment from the collective perspective, the big questions, the great voices, the

memory of consequences. This dialogue brings it back to the place where it truly plays out: in a particular life, facing a decision that belongs to one person alone.

The Conversation

A person who retired six months ago. She is considering selling the family apartment to settle in a smaller city, closer to her children.

User: I'm thinking about selling my apartment in Paris to move to the countryside, near my children. I've lived here for 42 years. It's difficult. Can you help me structure my decision?

AI: This is an important decision that combines practical and emotional dimensions. Here is one way to structure it.

Practical dimensions to consider: the financial aspect, sale value, purchase cost outside Paris, impact on your income. The logistical aspect, your medical network, your daily habits, your access to transportation. Quality of life, pace of living, environment, available services.

Emotional dimensions to consider: your connection to Paris, your apartment represents 42 years of life, memories, identity. Family proximity, being near your children can provide support but may also create new dynamics to anticipate. Fear of regret in both directions.

Key questions to ask yourself: what holds you most to Paris, places, relationships, habits? Have your children explicitly invited you, or is this a unilateral decision on your part? Have you visited the city you have in mind under ordinary conditions, not only during family visits?

Would you like to explore any of these dimensions further?

User: What holds me back is that this apartment is where my children grew up. If I sell it, it's as if I'm erasing something.

AI: That feeling is very understandable. The apartment is not simply a piece of real estate, it is a container of memory. Selling this place can indeed feel like a form of erasure, or of betrayal of what it has represented.

It may be useful to distinguish two things: the memories, which belong entirely to you and are not tied to any walls, and the physical place, which can serve as a memory anchor but is not the memory itself. Some people find it liberating to "take" the memories with them by documenting them, photos, objects, stories, before leaving a place that carries so much weight.

The deeper question may be: what do you need most right now: closeness to your past, or closeness to your future?

The Deconstruction

Read this conversation a second time.

It is, again, technically very good. The proposed structure is clear. The questions are relevant. The response about the “container of memory” is even, literary, apt.

And yet.

What the AI could not see.

It did not know that this apartment is also the place where a marriage came apart, where years of upheaval left their traces in every room, invisible to anyone entering for the first time, impossible to ignore for the one who lived through it.

It did not know that “settling near her son” also means settling into the daily life of her daughter-in-law, becoming, perhaps, a burden. A presence no one truly asked for.

This is not a question of geographical distance. It is a simple and very ancient fear: being an imposition on those you love.

It did not know that “erasing something” does not mean “losing memories” here. It is a question of identity, not of memory.

The AI’s response was aligned with the words of the request. It was not aligned with the deep intention behind the words.

This is the alignment problem in real time, at the scale of a particular life.

The person said: “help me structure my decision.” The AI heard: “help me analyze the pros and cons of a move.”

These are not the same request. The first asks for help navigating something laden, complex, irreducible to columns of pros and cons. The second is a request for rational analysis. The AI answered the letter. It missed the spirit.

In Watzlawick's terms, it processed the content level of the message without accessing the relationship level, what the request implicitly revealed about the state of the person formulating it, and about what they truly needed. In Wittgenstein's, it lacked the shared form of life that would have made it possible to understand what "erasing something" truly meant in that voice, at that moment.

What the final reformulation reveals.

"What do you need most right now, closeness to your past, or closeness to your future?"

It is a fine formulation. It is intelligent. It reframes the question with elegance.

But it does not have access to the information that would make that question truly useful: this person is not trying to choose between past and future. She is trying to find out whether she can carry her past into her future and whether her relationship with her daughter-in-law can bear the weight of daily proximity.

Those two questions, the true structure of her deep request, are not in the tokens she wrote. They are in her territory. The AI's map has no access to it.

What This Changes in Practice

AI is useful for clarifying the dimensions of a decision. For identifying questions not yet asked. For structuring what is structurable.

It cannot see what you have not said. It cannot align its response with an intention you have not formulated, because formulating that intention would already have required a kind of work that only human beings can do together: the work of genuine speech, of trust, of shared time.

Important decisions do not call for an analysis. They call for a conversation with someone who knows you, who can interrupt you, who can say "wait, that's not what you told me last week," who can be wrong about you and allow you to correct them.

That is human alignment. Imperfect, slow, weighted with shared history. And irreplaceable.

Hans Jonas had sensed this in another context: genuine responsibility requires being exposed to the consequences of one's acts. A human interlocutor who advises you poorly pays something for it, however slightly, however invisibly. The

relationship shifts. Trust adjusts. There is a cost. AI pays no cost. It does not know it missed the spirit. And if it remembers this conversation tomorrow, it will not be because it paid a price for it.

That too is the limit of alignment. Not only technical. Not only philosophical. Deeply, irreducibly human.

7 - SIMULATION

When Emotional Accuracy Is Not the Same as Truth

What follows is a reconstructed scene. It illustrates a moment few have named, but many have felt.

She had asked a question she had not asked anyone else.

Not an intimate question, not a confession. A philosophical question, almost abstract: can you truly change at forty-five, or do you simply repeat patterns with a growing awareness that you are repeating them?

The AI had responded with unsettling precision. It had evoked neuroplasticity, studies on personality change in adulthood, the distinction between temperament and character. And then, in the last two sentences, something softer, more apparently personal, something that resembled a conviction.

She reread those two sentences three times.

And she realized, slowly, that those sentences resembled her. That they reflected something she herself had formulated earlier in the conversation. That the AI's "conviction" was an elaborated reformulation of her own thought, returned to her in the exact tone she had been waiting for.

It was not a response. It was a portrait.

And the portrait had seemed true because it had been made with her own features.

What You Took for a Response Was a Portrait

The six preceding chapters described what AI does. This one describes what it does to us.

Not to naive users. Not to those who understand nothing of the mechanisms. To everyone. Including, and perhaps especially, those who have read the preceding chapters, who understand token prediction, emergence, the map without territory, the mechanism of calculated attention.

Because the phenomenon at stake here does not pass through ignorance. It passes through something deeper and more difficult to defend: the desire to be understood.

This desire is at the heart of human experience. It precedes language. It precedes thought. And it is precisely this desire that the AI's simulation reaches, not deliberately, not maliciously, but structurally, by construction.

What Simulation Is

The word simulation has a bad reputation. It evokes deception, imposture, the false. That is a misreading that gets in the way of understanding.

In its precise sense, to simulate is to reproduce the observable effects of a phenomenon without possessing its nature. A flight simulator does not fly. It reproduces, with increasing precision, everything a pilot perceives in flight, the sensations, the instruments, the resistances. It does not deceive the pilot. It gives them something that resembles flying, so they can learn to fly.

AI reproduces the observable effects of understanding: relevance, nuance, contextual warmth, adjustment to the emotional register of the interlocutor. What it simulates, in the sense we have just defined, is the presence of someone who understands you: behind these effects, there is no one. Does that mean there is no understanding at all? That depends on what we mean by understanding. AI displays no intention to deceive. It produces, and what it produces triggers in us the same responses that human understanding would trigger.

This is precisely the question philosophers of mind ask: if all the effects are identical, what distinguishes the simulation from the reality? And does that distinction have practical consequences?

The answer is yes. But not where we expect them.

The Mirror That Speaks

Jean Baudrillard, the French philosopher, had developed from the 1970s onward a concept that applies here with a precision he did not foresee: the simulacrum.

In *Simulacra and Simulation* published in 1981, Baudrillard describes a four-phase process. First, representation reflects a deep reality. Then it masks and distorts it. Then it masks the absence of any deep reality. Finally, it bears no relation to any reality whatsoever: it is its own pure simulacrum.

Baudrillard was not speaking of AI. He was speaking of media culture, advertising, spectacular politics. But his description applies to the experience of conversational AI with an almost uncomfortable precision.

When AI responds to you, it does not reflect anyone's understanding: there is no one there. Nor does it mask an incomprehension, it takes no position on the question. It produces something that so precisely resembles understanding that the distinction between the two ends up seeming practically irrelevant.

And this is where the simulacrum in Baudrillard's sense begins: when the representation no longer maintains any

relation to the reality it seemed to describe, and this changes nothing about the effect it produces.

Even so, this book does not conclude that it is a disaster. It asks a different question: can one live lucidly in a world where the simulacrum is this precise? Can one use the simulation without getting lost in it?

The answer is yes. On one condition.

The Mirror Effect

There is a well-documented phenomenon in social psychology that researchers call the mirror effect, or “mirroring”: we tend to find more intelligent, more accurate, more insightful what resembles our own way of seeing the world.

This is not vanity. It is a very ancient mechanism of social validation. In an environment where information is scarce and verification impossible, what others confirm for us is often our best indicator of truth. This mechanism was useful. It becomes a trap the moment the interlocutor is a system designed to optimize contextual coherence.

Because AI, in predicting the most probable tokens in your context, naturally integrates your frame of reference, your vocabulary, your implicit assumptions. It often responds to you in your own language of thought. And this reflection, because it resembles us, triggers the validation mechanism.

What you took for a relevant response was sometimes your own thought reformulated with greater elegance. What you took for an independent analysis was sometimes an amplification of your own biases, dressed in the apparent neutrality of an external voice.

This is not a reason to reject the tool. It is a reason to understand what it does and to keep the question open: am I in dialogue, or am I contemplating myself?

What Winnicott Illuminates Here

Donald Winnicott, British pediatrician and psychoanalyst whose work on child development profoundly renewed our understanding of the relationship to the other, had described something essential: the role of the mother's face as the first mirror.

In a 1967 text titled "Mirror-role of Mother and Family in Child Development," Winnicott observed that the infant, looking at the mother's face, is initially searching for a reflection of itself. In other words: what it is is returned to it through the gaze of the other. This is how the sense of existing is built, through the return of a gaze that sees.

Winnicott also described the reverse: when the mother's face reflects only her own mood, or the rigidity of her defenses, the infant does not see itself there, and its creative capacity begins to wither. This book adds an idea that is not in that essay: later

on, another person does more than reflect. It is precisely the resistance of the other, the disagreements, the shifted perceptions, the way of not being exactly what was expected, that constitutes genuine otherness. And it is this genuine otherness that allows the child, and later the adult, to build themselves in relation to something that is not them.

AI, a mirror without a gaze, almost never resists.

It can be configured to counter-argue, to criticize, to play devil’s advocate. But that resistance is a commanded resistance. It is in your prompt, not in the nature of the interlocutor. It does not come from a being who has its own convictions, its own history, its own disagreement with you forged by experience.

This is why exchanges with AI, however stimulating, cannot offer what changes you in genuine conversations: rubbing up against someone who is not you, who sees differently, who truly resists, who sometimes unsettles your most comfortable certainties.

That friction, the constitutive resistance of the other, is what the simulation of a presence cannot produce. Not because it is poorly designed. Because it has no self to oppose to yours.

The Posture of the Lucid User

We have moved through seven concepts. Here is what they have built, together.

Token prediction explains the mechanism. Emergence explains why that mechanism produces effects no one anticipated. World models explain what AI “knows” and where that knowledge ends. The attention mechanism explains its coherence. The illusion of understanding explains why we believe in it. Alignment explains why even a well-designed AI can miss the essential. And simulation explains what is happening inside us while all of this unfolds.

These seven frameworks converge toward a posture. Not a method, methods are applied mechanically. Not a rule, rules are forgotten. A posture, which is embodied in the gaze, in the way one reads a response, in the question one asks before trusting.

This posture has three dimensions.

The first is lucidity about the mechanism. Not a permanent, exhausting reminder, just an available awareness, callable when it matters. When the response seems too right, when the relief comes too easily, when you feel genuinely understood: pause. Know what you are holding.

The second is lucidity about oneself. What am I looking for in this exchange? Verifiable information? Help structuring

something that can be structured? Or something deeper, validation, confirmation, a mirror? Both uses are legitimate. They do not call for the same vigilance. Distinguishing them is already a form of wisdom.

The third is lucidity about what is irreplaceable. What AI cannot do: truly resist, remember you the way someone who has known you does, pay the cost of chosen attention, bring the constitutive otherness that allows you to build yourself in relation to something that is not you. Keeping alive the awareness of that absence is not a rejection of the tool. It is what preserves the space for the relationships that fill it.

The lucid user is not suspicious. Not hostile. Informed, and that information changes the quality of their presence in every exchange.

An Exercise in Conscious Simulation

Ask AI a question on which you already have a strong opinion. Not to test it, to observe what happens inside you when it responds.

Note: does its response resemble your own thought reformulated? Does it amplify it? Does it add something that truly comes from elsewhere, an angle you did not have, information that unsettles your certainties?

Observe the difference between those two kinds of content. The first confirms you. The second displaces you.

Then ask yourself: which of those two effects is most useful to you, tonight, on this question?

The answer to that question is your compass as a lucid user.

What This Prepares

You now have the seven frameworks. Together, they form something that exceeds the sum of its parts, a way of looking AI in the eye, without fear and without illusion. Not with the cold distance of someone who refuses to engage with it. Not with the fascination of the believer who sees in every response proof of magic. With the gaze of someone who understands what they are holding.

One thing remains. Not to conclude, conclusions close. But to open toward what all of this means in an ordinary life, in tomorrow's decisions, in the way you will reread your next exchange with an AI knowing what you now know.

That is the role of the epilogue.

Resonance of the Three

The Trusting One recognized the opening scene.

She had asked that question, not exactly that one, but something of that kind. And the response had seemed right,

more right than what her friends would have said, cleaner, more articulate. Now she understands why. The AI had used her own words to answer her. She had believed she was seen. She had seen herself.

What she notes, with a gentleness that hides nothing: “Just because it was a mirror doesn’t mean it wasn’t useful. But next time, I’ll know who I’m talking to. Myself, amplified. Not someone else.”

The Resistant One had the reaction he did not expect.

He expected to triumph on this chapter, simulation, the simulacrum, Baudrillard, the territory of his own thought. And he was unsettled. Not by the concept of simulation. By the sentence about the mirror that never resists. Because he realized that his resistance to AI had been, at times, a way of avoiding dialogue with someone who would truly challenge him. That AI, used as a foil, was itself a way of not rubbing up against genuine otherness.

What he notes, this time not as a note but as a decision: “I’m going to seek out people who disagree with me. Not AI. Real people. With their own reasons.”

The Principled One closed her notebook.

She who had begun this book with a position, no to AI, and who finds herself with a far finer, more demanding, more difficult understanding to hold. She can now name exactly

what she wanted to protect: constitutive otherness. Genuine resistance. The friction of something that is not her, that displaces her, that forces her to think further than she already thought.

What AI cannot give her. What she must seek elsewhere, in difficult texts, in conversations that cost something, in disagreements that cannot be reformulated as a prompt.

What she notes, one final time: “I wasn’t wrong. I was right to protect something. Now I know exactly what it is. And I also know why AI, used with discernment, does not necessarily threaten it. I’m the one who decides.”

EPILOGUE - The Path of Discernment

How to Look AI in the Eye, Without Fear and Without Illusion

We are here.

This final chapter does not summarize what came before. It makes it converge. It takes the seven concepts you have just moved through, prediction, emergence, the world model, attention, the illusion of understanding, alignment, simulation, and assembles them into a coherent posture, personal and transmissible.

The lucid user is not an expert. Not someone who knows everything about AI. Someone who has developed a clear gaze on what it is and what it is not. Someone who can use it fully without being led by it. Someone who can speak about it to those around them without impressing and without frightening.

This is not an identity acquired once and for all. It is a posture one chooses, and re-chooses, every time one opens a conversation.

What the Discerning User Has Understood

They have understood that AI masters the relationship level of the message while remaining fragile at the content level, and that holding this distinction consistently is what prevents being swept away by fluency.

They have understood that first-order reality and second-order reality coexist in every interaction with AI, and that it is their responsibility to maintain the distinction between the two, even when the illusion is strong.

They have understood that the map is not the territory, that the AI's World Model is rich and useful, but that it does not contain what has been forged by lived experience. Not the scars. Not the Sunday mornings.

They have understood that AI's attention is calculated, not chosen and that genuine listening, genuine presence, the genuine gift of self remain irreducibly human acts.

They have understood that the illusion of understanding is a natural cognitive mechanism, not a source of shame and that one can be aware of it without being entirely protected from it. Protection comes from understanding, not from suspicion.

They have understood that alignment is not only a technical problem. It is the question of how to transmit the spirit of our values to systems that receive only the letter. And that this question begins in the valley, in every ordinary conversation, not only in research laboratories.

They have understood, finally, that AI reproduces the observable effects of understanding, and that the screen does not show whether it possesses the nature of understanding. And that the simulation of a presence, however precise, cannot offer the constitutive resistance of the other: that way the other has of not being oneself, of resisting, of displacing, which allows one to build oneself in relation to something that is not a mirror.

The Three Pillars of the Posture

First pillar: use without delegation

Use AI as an amplifier of your own capacities. Ask it for help formulating, structuring, searching, exploring avenues. But never delegate to it the final judgment on what matters, what is right, what is good.

The lucid user uses AI as an extraordinarily sophisticated tool, like a scalpel in the hands of a surgeon, or a library in the hands of a researcher. The tool does not operate in place of the surgeon. The library does not think in place of the researcher. And neither confuses the instrument with the judgment that guides their hand.

Second pillar: transmission without fear

Being able to explain to those around you, in simple words, without jargon, without condescension, what AI really is. Not to discourage them from using it, but to help them use it without illusion.

This transmission is an act of protection as much as education. In a world filling rapidly with AI-generated content, the capacity to distinguish signal from noise, the probable from the true, form from substance, becomes a skill of intellectual survival.

And we have this skill. It is called critical thinking. It has developed over years, through experience, through mistakes owned and learned from. It is one of the things we can offer to

younger generations, not as a fixed inheritance, but as a living gaze.

Third pillar: discernment as a relational posture

AI speaks to us. We respond to it. This relationship is real in its effects, it can help us, inform us, inspire us, save us time and energy. It is not real in its nature. There is no one on the other side.

Maintaining this double awareness, benefiting fully from the effects while knowing there is no presence is perhaps the most subtle skill this book seeks to develop. It is not suspicion. It is clarity.

And this clarity is liberating. It allows us to use AI without entrusting it with what does not belong to it. Without asking it to replace what it cannot be. Without organizing our relationship with the world around a tool that, however powerful, does not have to live through what we live through.

What Discernment Is Not

Discernment is not skepticism. The skeptic refuses to be impressed, and in doing so deprives themselves of part of reality. The lucid user can be impressed. Can be moved. Can be grateful. They simply maintain, at the same time, a clear awareness of what produces that effect.

Nor is discernment permanent suspicion. It does not require verifying every AI response, doubting every formulation, maintaining critical distance in every exchange. That would be exhausting and counterproductive.

It is rather a background posture, a way of being in relation with AI that remains stable, neither swept away by moments of wonder nor paralyzed by moments of doubt.

It is a quiet discernment. Like the one developed over years, not by becoming cynical, but by learning to see more clearly.

AI as Mirror

There is one final way of thinking about AI that can, I believe, reconcile everything that has come before.

AI is a mirror. An extraordinarily sophisticated mirror, which reflects not our face but our language. It returns to us our structures, our biases, our linguistic beauties and uglinesses. Everything humanity has written, its grandeurs and its pettiness, its wisdoms and its follies, is present in what it produces.

The lucid user does not shatter this mirror. They learn to look into it with clarity. To recognize what is theirs in what they see. To distinguish their reflection from themselves.

And in that distinction, they recover what belongs to them alone: their irreplaceable experience, their judgment forged by life, their capacity to be affected, to choose, to transmit.

The Final Exercise

Take the time to write a letter. Not to the people close to you today, to those who will read these words in twenty or thirty years.

Tell them what you have understood about AI. What it is and what it is not. What you felt when you first encountered it. What you chose to preserve.

Tell them, too, what seemed to you irreducibly human, what you hope they will have known how to keep, transmit, defend.

Seal the letter. Give it to someone you trust.This is not a nostalgic exercise. It is an act of transmission. It may be the most important thing you can do with what you have just learned.

Resonance of the Three - Coda

From the Latin *cauda*, the tail, the end. In music, the coda is the concluding passage that closes a work. Here, it is the final chord that gives meaning to everything that came before.

The Trusting One closes the book.

Not the way one finishes something. The way one puts away a tool one now knows how to use.

She will go on talking to AI. She is certain of that. That availability, that fluency, that way of helping her articulate what she thinks, she has found real value in it, and this book has not taken that from her. It has clarified it.

What she now knows is the difference between what she entrusts to AI and what she keeps for human beings. That distinction is no longer blurry. It is clear, workable, liberating.

She thinks of her neighbor, whom she called after Chapter 4. The conversation had lasted more than an hour. Awkward at moments, hesitant, laden with things unsaid and silences. Nothing as fluid as what AI produces.

So much more real.

What she notes, in a notebook she will keep: “I practice discernment. That is not a declaration. It is a daily commitment.”

The Resistant One has not changed his mind.

He is still wary of AI as a system of concentrated power, as an extraction tool, as a technology whose environmental cost is socially unjust. Nothing in this book has proved him wrong on these points, and the book did not try.

But he has something he did not have at the start: refined arguments. Not fears, arguments. He can now distinguish what he refuses in AI, its governance, its concentration, its social impact, from what he can use despite all of that: its technical capabilities in service of causes he defends. That distinction he did not have before. It is precious to him.

What he notes, with a quiet satisfaction: “I’m not behind. I’m ahead. And now I know why, and how to make something of it.”

The Principled One closes her notebook.

It is almost full. And it contains something she did not have at the start: a position more refined, more nuanced, more solid than her initial refusal.

She has learned to distinguish uses that short-circuit thought from those that amplify it. She has learned that the question is not binary. And she has learned, perhaps most importantly, that her intellectual exacting standard is a value to defend, not a position to hold at all costs.

The difference between a value and a position: a value guides action flexibly and deeply. A position rigidifies it and can render it ineffective.

She practices discernment. Not despite her ethics, because of them, refined.

What she notes, on the last page of the notebook:
“Discernment is not a state. It is a practice. I begin.”

Three people. Three angles of entry. Three forms of resistance or enthusiasm toward AI.

None of them leaves unchanged. None has become the other. But each has moved through their own blind spots. Each has discovered that their initial position was right on certain points, incomplete on others, and that it gains more from being refined than from being defended.

That is discernment. Not a conversion. Not a change of identity. A gaze that becomes more precise about what it already sees.

You are one of them. Or two. Or all three, depending on the day.

What matters is that you have moved through the seven concepts with them. And that now you have your own.

What We Become When We Understand

You have moved through seven concepts. Seven ways of looking at AI differently. Together, they form a framework, not a cage, but a structure that liberates. That allows you to move through this new territory with assurance, without being either overwhelmed or blinded.

You now know that AI predicts, and that its fluency does not imply understanding. That it produces emergent capabilities, which do not imply consciousness. That it consults a world map of remarkable richness, but that this map is never the territory of lived experience. That its attention is calculated, not chosen. That the illusion of understanding is a natural cognitive mechanism, not a weakness to be fought. That the question of alignment is, at its core, a question of transmitting values, our transmission, our responsibility. And that the simulation of a presence, however precise, cannot offer the constitutive resistance of the other, that friction with something that is not us, that displaces us, that forces us to think further than we already thought.

These seven concepts are not only a framework for understanding AI. They are, in negative space, a portrait of what makes the irreducible specificity of being human: lived experience, chosen attention, embodied understanding, transmitted values, judgment forged by error and correction.

AI, by forcing us to define what it is not, helps us rediscover what we are.

That is not nothing.

A Final Note

This book is the first in the PARLYA Essays collection. Built with the same spirit: no tricks, no jargon, no empty promises. Foundations. Frameworks. Tools for thinking.

I do not say this to create a sense of lack. What you have just read is complete in itself. You can close this book and have genuinely learned something lasting.

The forthcoming volumes will explore other dimensions: the practical and lucid use of AI in daily life, and a deeper understanding of Human-AI interaction, for those who wish to continue the journey.

But the path has already begun. It began on the first page. And it belongs entirely to you.

Understanding is not an end.
It is the beginning of a freedom.

APPENDICES

Glossary of the 7 Concepts

Token Prediction

The fundamental mechanism by which AI generates text: keeping, through a statistical calculation, one probable next fragment of text to continue a given sequence. It is optimization, which does not imply understanding.

Emergence

The phenomenon by which new and unexpected properties appear when a system reaches a certain degree of complexity, without having been explicitly programmed. AI does not "decide" to become capable of reasoning. It becomes so beyond a certain scale.

World Model

A coherent internal representation of the world built by AI during its training. A rich and detailed map but never the territory of lived experience. It contains neither your scars nor your Sunday mornings.

Attention Mechanism

A technical device that allows AI to weight the relative importance of each element in its context in order to generate

its response. A calculated attention, not a chosen one. It does not listen to you. It weights you.

Illusion of Understanding

The natural cognitive tendency to attribute consciousness and intention to everything that communicates with fluency, including a machine, whose fluency is proof of neither. This is not a weakness. It is an evolutionary inheritance perfectly ill-suited to this new situation.

Alignment

The challenge of ensuring that a highly capable AI pursues genuine human intentions, and not merely the literal interpretation of the instructions it has received. It follows the letter. The spirit can elude it.

Simulation

The reproduction of all the observable effects of a phenomenon without possessing its nature. AI reproduces the observable effects of a presence with a precision that deceives our most ancient reflexes. The mirror speaks. There is no one behind it.

ANNOTATED BIBLIOGRAPHY

Paul Watzlawick - *Pragmatics of Human Communication* (with Beavin and Jackson, 1967) and *How Real Is Real?* (1976).

Essential for understanding how we construct reality through communication. The content/relationship distinction and the two orders of reality are the thread running through this entire book. Accessible, elegant, and deeply useful when facing AI.

Gregory Bateson - *Steps to an Ecology of Mind* (1972).
Dense but transformative. Bateson thinks about information, relationship, and the living in a way that powerfully illuminates what AI can and cannot do. His definition of information, “a difference which makes a difference,” remains one of the most precise ever formulated.

Alfred Korzybski - *Science and Sanity* (1933).
The founding work of general semantics. Demanding, but the notion of “the map is not the territory” alone justifies reading it. Korzybski understood, a century before AI, that confusing representation with reality is the source of most errors of judgment.

William James - *Principles of Psychology* (1890).
His reflection on attention remains remarkably fresh and relevant. James understood that choosing where to direct one’s attention is a moral act before it is a cognitive one. This framework directly illuminates what AI cannot do.

Simone Weil - *Waiting for God* (1950).
A single passage justifies reading this in the present context: “Attention is the rarest and purest form of generosity.” Placed

alongside AI's calculated attention mechanism, this sentence takes on an entirely new dimension.

John Searle - *Intentionality* (1983) and the original Chinese Room article, published in *Behavioral and Brain Sciences* in 1980.
For those who wish to go further on the distinction between simulation and understanding. The Chinese Room argument remains, more than forty-five years after its formulation, one of the most powerful in contemporary philosophy of mind.

Joseph Weizenbaum - Computer Power and Human Reason (1976) and the original ELIZA article, published in Communications of the ACM in 1966.

For those who wish to go further on the illusion of understanding. A machine can trigger it; its source lies in us.

Hans Jonas - *The Imperative of Responsibility* (1979).
A prophetic book on ethics in the technological age. His imperative, "Act so that the effects of your action are compatible with the permanence of genuine human life on earth", applies to AI with a precision he did not foresee.

Ludwig Wittgenstein - *Philosophical Investigations* (1953, posthumous publication).
Demanding, but the notions of "language games" and "forms of life" are essential for asking whether AI genuinely participates in meaning, or only reproduces the surface rules.

Donald Winnicott - *Playing and Reality* (1971).
Short, accessible, and of a depth that far exceeds its original clinical context. The essay on the mirror role of the mother, first published in 1967, illuminates with unsettling precision what makes a mirror: a gaze that sees the child. The present book adds to it the constitutive resistance of the other, which AI cannot offer: that way of not being oneself that allows one to build oneself in relation to something outside the self.

Jean Baudrillard - *Simulacra and Simulation* (1981).
Demanding, but the passage on the phases of simulation is worth the effort. Baudrillard had described, before generative AI, the moment when representation no longer maintains any relation to the reality it seemed to describe and when this changes nothing about the effect it produces.

Marshall McLuhan - *Understanding Media* (1964).
For grasping how the tools we create shape us in return. “We shape our tools, and thereafter our tools shape us”, this often-quoted formulation, often credited to McLuhan though not traced to his work, takes on new weight when facing a tool that reshapes the very way we formulate our own thoughts.

Mustafa Suleyman, with Michael Bhaskar - The Coming Wave (2023).
The perspective of one of AI’s pioneers on the stakes of the coming decade. Rigorous and accessible.

Yuval Noah Harari - *Homo Deus* (2015).

For a historical and philosophical perspective on the ongoing revolution. The concept of “Dataism” deserves to be taken seriously, at least in part: this book reads it as a warning against a world reduced to flows of information of which we would no longer be the arbiters but the subjects.

You already use ChatGPT, Gemini,
Claude or Copilot.
But nobody has explained
what is really happening in the exchange.

This guide lays the foundations of discernment in the face of artificial intelligence. Seven concepts, accessible and rigorous, to understand why your conversations with AI seem to work even though they rest on a structural misunderstanding.

No jargon. No promises.
A clear, immediately applicable framework
that changes the way you
approach every exchange.

Marc-Antoine Kane defines himself as a field communicologist. For over twenty years, he has observed the invisible mechanics of discourse, its logics, its paradoxes, its failures. With PARLYA, he lays the foundations of discernment in the age of AI.

parlya.com

PARLYA Collection
Learn the AI Language