CONES OF MEANING

Meta-Rationality, Outfinitism, and Alignment in the Age of Generative Intelligence

A companion volume to Outfinitism:
Meta-Rationality and the Limits of Every Frame

CONES OF MEANING

Meta-Rationality, Outfinitism, and Alignment in the Age of Generative Intelligence

A companion volume to Outfinitism: Meta-Rationality and the Limits of Every Frame

2026

"We never see the world without a filter. Wisdom begins when we also see the filter."

Author's Note

This work was created under the umbrella of Axiologic Research as part of two interconnected research programs, one dedicated to Artificial Intelligence and the other to Outfinitism, both led by Sînică Alboaie. To explore the details of how these two programs intersect and build upon each other, we invite readers to visit the Outfinity Venture Studio page at www.outfinity.ch.

While Artificial Intelligence language models were used to assist with text generation, the foundational philosophy, thematic structure, and analytical approach derive exclusively from the author's decades of dedicated study of social phenomena, social technologies, and genuine hard technologies resulting from various European and self funded research projects. Recently, within the Achilles research project, we have been deeply studying AI generated literature and its automated verification using neuro symbolic approaches; all the free books made available to the public are part of the suite of works we use to test our latest technologies.

Note on This Volume

This volume continues and complements *Outfinitism: Meta-Rationality and the Limits of Every Frame*. Its central terms are reintroduced wherever the narrative requires them, so that the text can be read independently, while their meaning remains connected to the philosophical project developed in the earlier volume.

Meta-rationality refers here to the practice of reasoning rigorously while also examining the frame, purpose, and distribution of authority that make reasoning possible. Outfinitism treats the actual limits of models, evidence, computation, institutions, and bodies as frontiers that must be declared, judged, and, where possible, pushed through finite and revisable attempts. Neither term is presented as an already established discipline or as a universal stage of psychological development.

The expression "alignment as a state of consciousness" is used only as a working metaphor. The book does not attribute phenomenal experience, a persistent self, or moral authority to language models; it discusses a functional state of orientation, observable in the way a system selects relevance, handles uncertainty, and orders values in context.

Artificial intelligence systems were used as editorial and research instruments for elaboration, criticism, and reformulation. The coherence of the text is not evidence of novelty, correctness, or empirical validation. The selection of the thesis and authorial responsibility belong to Sînică Alboaie.

Contents

PREFACE

The Second Volume: From the Limits of Every Frame to the World Produced Through Frames

This book continues a question opened in *Outfinitism: Meta-Rationality and the Limits of Every Frame*. The starting point there was simple and uncomfortable: reason never operates in a vacuum. Every act of knowing begins within a frame that establishes what exists for analysis, which differences matter, what can be measured, what purpose inquiry or decision is meant to serve, and who receives the authority to turn a conclusion into action. Meta-rationality was the name given to a maturity capable of examining that frame without abandoning rigor. Outfinitism was the concentration of that maturity on limits: the limits of the model, the evidence, computation, the institution, the body, and power. It did not propose a final system, but a discipline of finite attempts open to correction and extension. [ALBOAIE-2026]

The present volume does not simply repeat that argument. It moves it into an environment in which the problem of frames becomes more visible and more urgent. Generative models can produce, in a very short time, texts, explanations, hypotheses, programs, images, scenarios, and justifications that until recently required substantial human investment. Yet they do not generate from a neutral space. Every answer is oriented by training data, architecture, post-training, the prompt, the conversational context, connected tools, and the criteria by which the system was taught to appear useful. Generative intelligence has not eliminated frames. It has made them productive at an unprecedented scale.

This is where the central metaphor of the book appears: the **cone of meaning**. A culture, an institution, a person, or a model does not simply see a collection of facts. It orients itself within a field of possibilities, draws some meanings closer, pushes others away, recognizes certain analogies, and ignores other relations. A cone is neither the world nor a complete theory of mind. It is an image for the selective direction through which a plurality of elements becomes a coherent perspective. The metaphor will be connected, carefully, to the geometry of vector representations used in language models, but it will not be presented as proof that meaning is literally a geometric object. It is useful precisely because it preserves two intuitions at once: orientation produces coherence, and in a high-dimensional space many partially overlapping orientations can exist.

The image also helps us understand historical epochs. An epoch is not a person with a single belief, but it can develop a dominant repertoire of questions, metaphors, institutions, and criteria of legitimacy. Modernity has looked strongly through the cones of rationalization, measurement, general law, progress, and science. These cones have produced extraordinary achievements. They have reduced the arbitrariness of rank, made claims contestable, created technologies capable of transforming life, and offered legal protection against personal power. This book does not begin from a revolt against those achievements. It begins from the problem of their success. A grid that solves major problems acquires prestige, and prestige expands its jurisdiction. Method becomes a total vision, the indicator becomes reality, procedure becomes justice, prediction becomes permission, and science can be transformed into scientism precisely by those who borrow its authority without accepting the discipline of its limits.

Meta-rationality and outfinitism are introduced here as two complementary movements. The first opens the field: it makes the frame visible, compares perspectives, asks what has been excluded, and searches for a more adequate formulation. The second returns analysis to the condition of the finite agent: time runs out, data are partial, computation costs, institutions have mandates, decisions create losses, and some errors cannot be repaired. In a deliberately approximate analogy, they can be seen as a yin and a yang of contemporary wisdom. Meta-rationality prevents rigidification within a single map; outfinitism prevents the unlimited proliferation of maps and turns openness into responsible decision. Each contains the seed of the other. Comparing frames is itself finite, and a limit sometimes cannot be understood without changing the frame in which it was defined.

The theme becomes especially important when a language model seems to understand a philosophy it has not encountered in this form. The experience that motivated this book was precisely of that kind. A few indications about meta-rationality and outfinitism were enough for a model to produce equivalent formulations, analogies, and consequences that seemed to preserve the direction of the human intention. We cannot prove that the model had never encountered related fragments. We can, however, explain why it did not need to have memorized a treatise with the same definitions. A large language model learns distributed relations among contexts, words, roles, and argumentative patterns. A prompt can temporarily orient this semantic ecology and make a family of coherent continuations more probable. The sense of understanding appears when the system does not merely repeat the words, but preserves the relations among limits, pluralism, decision, verification, and wisdom.

This performance is real, but its status remains open. It may represent semantic generalization and functional understanding without phenomenal experience or a self

that possesses the perspective. It may also be a fragile simulation, dependent on context. The book will not decide the metaphysical question of artificial consciousness. It will nevertheless use the risky formula "alignment as a state of consciousness" to illuminate a question that strictly behavioral vocabulary tends to simplify: what global orientation makes certain interpretations, levels of certainty, and value priorities more probable before an answer is formulated?

The book therefore follows two lines of inquiry that eventually meet. The first is cultural. Major ideas escape their original domains and become metaphors, memes, institutions, and ways of seeing. Relativity, evolution, quantum theory, cybernetics, information, legal rationality, and science have all had cultural lives that exceeded their technical content. Some transfers were fertile; others preserved prestige while losing constraint. The second line is technical. LLMs show that meaning can be operated through distributed representations, contextual attention, and probabilities of continuation. They can rapidly carry different cones and produce coherent textual worlds from within them. The meeting of the two lines forces us to rethink the relation between epistemology and axiology: not only what can be generated and justified, but what is worth generating, for whom, at what cost, and under whose authority.

This volume is therefore complementary to the first. *Outfinitism* described the limit as an object of wisdom and proposed the open finite as an image of revisable attempts. *Cones of Meaning* asks how orientations are formed that make limits visible or invisible, how they circulate through culture, and how they can be imitated or amplified by artificial intelligence. The first volume looked at the vessel and the water; this one looks at the direction of the light through which we see the vessel, the water, and sometimes only our own reflection.

The metaphor must not be absolutized. Sometimes the network, the field, the ecosystem, the role, or the practice describes the phenomenon better. A meta-rational book must allow its own concepts to be replaced, and an outfinitist one must acknowledge that it offers a finite compression of a larger problem. Its ambition is not to close reality inside a new doctrine, but to make visible relations that can improve the way human beings and artificial systems generate, choose, and act.

The Age of Cheap Coherence

Much of modern culture was organized around the scarcity of production. A book required time, serious analysis required training, a working program required experience, and a convincing image required an author or team capable of executing it. These barriers guaranteed neither truth nor value. They excluded voices, preserved hierarchies, and allowed institutions to confuse prestige with merit. Yet the cost of production did have a filtering effect. Before an idea could occupy public space, someone usually had to invest time, reputation, money, or labor. Generative intelligence radically lowers that cost. Verbal coherence, illustration, variation, summary, and justification become available almost instantly.

The change can be stated without dramatization: the cost of expression is falling, while the cost of judgment is becoming more visible. Producing a plausible version is no longer as difficult; determining which version deserves attention, verification, and implementation becomes harder. A person can request ten strategies in a minute, but the time needed to understand their consequences does not fall in the same proportion. A model can generate bibliographies, code, and arguments; references must still be checked, programs tested against the real purpose, and arguments compared with evidence and consequences.

This is an epistemic crisis because it becomes difficult to separate what is true, probable, derived, invented, or merely well phrased. It is also an axiological crisis because no purely technical function can decide what deserves to be produced. A hundred correct texts can consume the attention needed for one important text. A system can optimize for the user's immediate usefulness while degrading the shared information environment. It can generate personalized explanations that increase understanding or personalized justifications that enclose a person within existing beliefs. Abundance does not free us from value; it forces us to make value more explicit.

The crisis arrives in a culture already marked by an unresolved tension around reason and science. On the one hand, modern science, law, engineering, and administration produced real emancipation. They made claims criticizable, reduced dependence on inherited authority, and built institutions capable of learning. Any philosophy that treats them only as expressions of one culture among others becomes unserious. On the other hand, their success produced the temptation to transform a set of methods into an implicit metaphysics. What cannot be measured becomes secondary; what does not enter the model appears unreal; what can be predicted is treated as legitimate to administer;

and the phrase “science says” sometimes carries moral and political decisions that no experiment has demonstrated.

The distinction between science and scientism is essential. Science is a family of practices through which hypotheses are exposed to evidence, criticism, replication, modeling, and revision. Scientism appears when the prestige of those practices is extended to questions that the method cannot decide by itself, or when one interpretation is presented as the necessary voice of “science.” Science can estimate the effects of a policy; without normative premises, it cannot decide which distribution of costs is just. It can describe correlates of well-being; it cannot reduce the meaning of a life to the indicators available. It can show that some beliefs about the world are false; it cannot replace, with a single score, all relations among truth, freedom, dignity, loyalty, risk, and hope.

Scientism is not too much science, but too little memory of its limits. It simplifies the real history of knowledge, in which models have domains, instruments have resolutions, data are selected, and scientific communities are human institutions exposed to incentives and power. Yet the critique of scientism becomes dangerous when it slips into anti-science. The perspectival character of models does not make all claims equal. The presence of values in research does not turn experiment into opinion. The possibility of institutional error does not transform personal intuition into a substitute for evidence. Meta-rationality and outfinitism must preserve this balance: defend the discipline of knowledge against arbitrariness while also defending reality against the claim that one discipline can exhaust it.

In the companion volume, rationality was described as reasoning within a frame, and meta-rationality as the possibility of reasoning about the frame as well. The formulation deserves to appear near the beginning because the story of this book depends on it. A frame is not merely an opinion. It contains a practical ontology - what kinds of things exist for the decision - a selection of relevance, a purpose, a scale, a time horizon, a distribution of authority, and a rule for accepting evidence. A well-formulated problem is already a small world. Reason can operate impeccably inside it and still produce the wrong conclusion for the larger world because the question excluded, from the beginning, what mattered.

A hospital that tracks waiting time captures the tension in a simple form. The indicator is legitimate, but once it becomes a target it can reorganize behavior: difficult cases are reclassified, consultations are shortened, and care that does not fit the number begins to look like inefficiency. The system has not abandoned rationality; it has become more rigorous within a frame that is too narrow. [GOODHART-1975] The pattern recurs wherever the proxy is rewarded more directly than the purpose.

These examples do not ask us to eliminate metrics. Without measurement, good intentions can conceal failure. They ask us to preserve a distance between the indicator and the value it represents. Meta-rationality makes the choice of indicator visible and constructs alternatives. Outfinitism requires the claim to remain proportional to the actual coverage of the measurement and prevents boundary effects from being hidden. Axiology asks why that value was chosen, who gains from the optimization, and who bears what the indicator forgets.

This is where "cones of meaning" enter the scene. When we view a situation through the cone of efficiency, time, cost, yield, and bottlenecks become visible. Through the legal cone appear competence, rule, precedent, and avenues of appeal. Through the medical cone appear diagnosis, prognosis, and the balance of benefit and risk. Through the cone of lived experience appear fear, dignity, relationship, and meaning. No cone is useless. The problem begins when one of them forgets its domain and becomes the only reality admitted. Meta-rationality does not require every situation to be viewed through every possible cone. That would be another impossibility. It requires the capacity to notice when an important cone is missing and to justify why a particular frame receives priority here and now.

Outfinitism adds what pluralism risks forgetting: comparing perspectives, costs, and action cannot wait for every dispute to be settled. Some decisions are irreversible; some limits can be overcome, others must be accepted, and still others are created artificially to protect an institution. The outfinitist question is what kind of limit we face, for whom, relative to what purpose, and with what resources. The open finite names a succession of real attempts, each limited yet capable of preparing the next extension. [ALBOAIE-2026]

Artificial intelligence concentrates all these problems. An LLM is at once a compression of cultural traces, an instrument of generation, a participant in dialogue, and a component of institutions capable of acting. It can move between cones faster than human beings can explain their limits. It can speak legally, therapeutically, scientifically, religiously, managerially, or poetically. This mobility creates the impression of an intelligence that sees above frames, but the model is not necessarily a neutral arbiter among them. It can be oriented by the prompt, by the interlocutor's style, and by preferences learned during post-training. It can produce a meta-rational critique and, in the next conversation, an equally fluent dogmatic justification.

For this reason, alignment cannot be reduced to asking whether an answer contains a prohibited claim. The system must interpret the situation, choose which information matters, distinguish a factual question from a normative one, manage conflict between usefulness and risk, and decide when to refuse, verify, abstain, or offer more than one frame. Contemporary alignment methods already try to induce such dispositions

through human feedback, constitutional principles, deliberation over policies, and supervision of the process. [OUYANG-2022] [BAI-2022] [GUAN-2024] [LIGHTMAN-2023] The hypothesis of this book is that we can better understand what these methods are trying to achieve if we speak of a **state of orientation**: not a fixed metaphysical property, but a contextual tendency to see problems through a particular set of relations, values, and limits.

The phrase "state of consciousness" will be used as a test of the metaphor, not as a conclusion about the ontology of models. It is useful only insofar as it can be translated into functional changes of salience, interpretation, and action; it becomes misleading when it confuses behavior with experience or shifts responsibility from human institutions to a presumed personality of the machine.

The story that follows begins with modern rationality and with ideas that exceeded their original domains. It then examines cultural cones, memes, and epistemic epochs. At the center of the volume, meta-rationality and outfinitism are reconstructed as two complementary movements, explained sufficiently for a reader who has not read the earlier book. The technical mirror follows: why a language model can seem to understand a newly formulated philosophy and what this tells us about the structure and fragility of meaning. From that mirror emerges the axiological problem of abundant generation and, finally, the hypothesis of alignment as orientation.

The book does not end with a grand architecture. There is no evidence that a single "outfinitist filter" would solve alignment. There are, however, reasons to prefer certain dispositions over performative certainty and rigidity: making the frame explicit, separating evidence from value, declaring the limit, formulating a relevant alternative, and preserving a path of correction. These dispositions can inform education, governance, and the evaluation of AI systems without being confused with a total solution.

The final test is applied to the thesis itself. Cones can become apophenia; pluralism can become relativism; the critique of scientism can become anti-science; outfinitism can become an excuse for weak standards; understanding hypocrisy can become a justification for manipulation; and meta-rationality can become a language of superiority for those who have the power to define which perspectives are admitted. A philosophy that studies limits is not credible if it hides its own. This is the condition under which the book should be read: as a possible counterweight, not as a new center of the world.

CHAPTER 1

REASON WITHIN THE FRAME

A Moral Technology That Sometimes Forgot Its Domain

The critique of rationality is easy when it forgets the world that rationality helped to make disappear. In societies organized mainly through rank, tradition, and personal authority, the question "why?" could receive a simple answer: because the person above has decided so. The rationalization of law, administration, and knowledge introduced a different requirement. A claim had to be supported, a rule had to be formulated, a procedure had to be known, and a decision could be compared with other decisions. No institution ever fulfilled these ideals perfectly, but their emergence changed the nature of legitimacy. Power could no longer rely exclusively on origin, prestige, or revelation; it had to produce reasons that could, at least in principle, be discussed publicly.

This is one of the great moral technologies of modernity. Science separates, imperfectly but essentially, the truth of a claim from the status of the person who makes it. General law tries to separate the treatment of a person from the whim of an official. Engineering makes the consequences of decisions calculable before they appear in the collapse of a structure. Medicine compares treatments and turns accumulated experience into criticizable evidence. Democracy, when it works, requires the reasons for exercising power to be stated in a language open to challenge. The alternative to these practices is not automatically the organic wisdom of the community. It may be superstition, obedience, impulse, corruption, or violence covered by mythology.

Meta-rationality therefore does not begin by weakening reason. It begins by protecting reason from its own uncontrolled extension. A method that is excellent within one domain can become dangerous precisely because its success makes the domain invisible. When calculation solves spectacular problems, the temptation arises to assume that every important problem must, in its essence, be a problem of calculation. When legal procedure reduces arbitrariness, the temptation arises to identify legality with justice. When experiment produces robust knowledge, the temptation arises to assume that what cannot be investigated in the same way cannot be known or does not matter. The instrument loses its edge and begins to resemble the world itself.

Understanding the frame helps us see how this error arises without turning rationality into an enemy. Every problem formulated for analysis already contains a selection. The relevant entities, variables, scale, time horizon, purpose, and criterion of success have been chosen. A map is intelligent through what it omits. A road map that

included every smell, conversation, stone, memory, and possibility of a city would not be a more complete map; it would become unusable. Models produce cognitive power through compression. The price is loss. The mature question is not whether we can construct a representation without loss, but whether the losses are acceptable for the concrete purpose and consequences.

A financial model may ignore the emotional history of a family when calculating a mortgage payment. The omission is justified if the question is strictly accounting. It becomes insufficient if the same model decides the family's eviction and claims to have exhausted the moral and political problem. A medical statistic can be correct for the population studied and still fail to decide, by itself, what risk is worth accepting for a patient who assigns different weights to longevity, pain, and autonomy. An administration can apply the same rule to everyone and still produce injustice if the data used to classify people have eliminated the relevant difference. Consistency is an important property. It is not synonymous with neutrality.

Herbert Simon made the finitude of the agent central to decision theory. Human beings do not search for the optimum in a fully known world; they simplify, use heuristics, and stop at solutions that are good enough under the available time and information. [SIMON-1955] This is not a shameful departure from intelligence. It is the environment in which intelligence exists. More computational power can extend the search, but it does not automatically choose the goal. More data can improve precision, but data are collected through categories. More sensors can record behavior, but they can make less visible what does not appear in the captured form. A society that measures everything may become a society that confuses the real with what its infrastructure knows how to count.

Digitization turns this confusion from an intellectual habit into an executable mechanism. A verbal assumption can be challenged in a meeting. An assumption encoded in a database, an eligibility rule, or a classification model can be applied millions of times before anyone recognizes it as a choice. An administrative category begins by describing and ends by organizing. People alter their behavior to fit the criterion. Institutions design workflows around the indicator. When a complaint appears, it must be translated into the language of the system that produced the problem. Formalization can offer predictability and, at the same time, reduce the possibility of appeal.

Here we can see why rationality cannot be separated from purpose. Logic can tell us whether a conclusion follows from premises. It cannot supply the value-laden premises by itself. If the objective of a platform is time spent, mathematics can optimize recommendation. It cannot decide that occupying attention is a legitimate purpose or that anger is an acceptable cost. If the objective of a hospital is to reduce expenditure,

analysis can compare treatments. Without a view of dignity, it cannot decide which forms of care are dispensable. If the objective of an economy is aggregate production, models can identify policies that increase it. By their technical form alone, they cannot declare unpaid care work, the distribution of risk, or ecological continuity unimportant.

This does not mean that values are immune to facts. A purpose can be incoherent, impossible, or based on false beliefs. A well-intentioned policy can produce the opposite effect. A value that refuses every contact with consequences becomes a slogan. Yet the fact that reality constrains purposes does not mean that it chooses among all possible purposes. Engineering can describe the costs of a bridge that serves traffic and of one that protects a neighborhood. It cannot decide by itself what kind of city is worth building. Calling this choice "political" does not make it irrational; it shows that instrumental rationality needs a rationality of ends and judgment about values.

The companion volume called meta-rationality the capacity to separate and then recombine these levels: observation, inference, purpose, choice, and the power that turns choice into consequence. [ALBOAIE-2026] The distinction matters because public conflicts often mix them together. One side says that "the data are clear" when the real dispute concerns who bears the cost. Another invokes identity or value to protect a factual claim that has been refuted. A report introduces a moral decision into a parameter, and the output is later presented as if morality had been eliminated through calculation. Meta-rationality does not claim that the levels can be perfectly isolated. Values direct attention, attention selects evidence, evidence can change values, and institutions decide what information comes to exist. The purpose is to make the transitions visible.

This visibility is also a problem of power. The person or institution that defines the frame partly decides who appears in the problem, which suffering becomes measurable, and which outcome can be called success. The greater the consequences and the weaker the possibility of challenge, the less legitimate it is for the frame to remain implicit. A formula is no more neutral than the question for which it was built. A model is no more objective than the relation among its data, its purpose, and the world over which it is authorized to act. Internal rigor can coexist with external injustice, and this possibility is not a peripheral accident. It defines the limit of rationality within the frame.

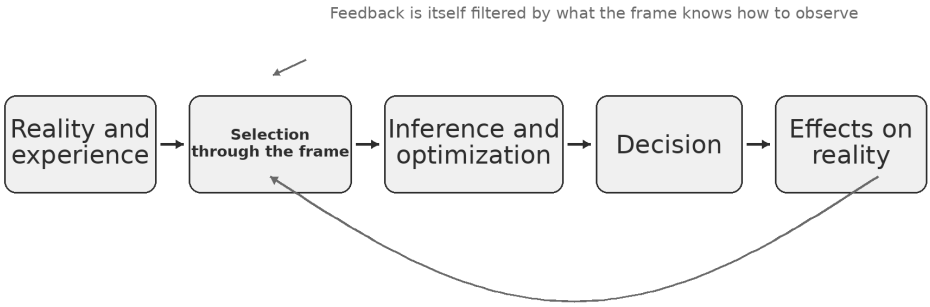

Diagram 1. The frame as an intermediary between reality and decision

From Precision to Scientism, Metrics, and Manipulation

If local success tends to expand the jurisdiction of a method, then modernity must be understood not only through its ideas, but through the dynamics of their exaggeration. Rationality emerged as a correction to arbitrariness and produced institutions capable of self-correction. Then, in some contexts, the ideal of rationalization was reduced to what could be formalized, standardized, and optimized. Max Weber observed the power and ambivalence of this transformation: the calculability of law and administration supports predictability, but it can also enclose life within a "casing" of procedures in which substantive meaning becomes increasingly difficult to formulate. [WEBER-1922] The rule protects against whim, but it can also become the language through which institutional whim declares itself impersonal.

Science faces a similar tension. Objectivity is not a pure view without an observer. It is a historical family of practices that reduce the dependence of a conclusion on the preference of a single observer: calibrated instruments, description of methods, replication, criticism, statistics, archives, and communities that can verify. [DASTON-GALISON-2007] [LONGINO-1990] The power of science comes precisely from exposing a hypothesis to a reality that may refuse to cooperate with the researcher's desire. But a scientific result does not enter society without interpretation. An average effect must be related to a person; a model must be extrapolated; one risk must be compared with another; and a decision must include costs the experiment did not measure.

Scientism appears when this mediation is forgotten. It is neither strong respect for evidence nor the ambition to explain. It is the extension of scientific prestige across the boundary between description and value, between model and reality, between correlation and authorization to intervene. The proposition "science has shown that X produces Y under certain conditions" may be solid. The proposition "therefore society must choose

Z" contains additional premises about priorities, rights, distribution of costs, and acceptable uncertainty. Sometimes those premises are justified. The problem begins when they are hidden in the shadow of the white coat or the graph.

Science	Scientism
Ties its claims to methods, domains, and the possibility of correction.	Turns the prestige of the method into a general authority over meaning and value.
Distinguishes data, model, interpretation, and decision.	Compresses these levels into the formula "science says."
Treats limits as part of the result.	Treats limits as details that weaken the message.
Can cooperate with ethics, law, and lived experience.	Treats them as noise if they cannot be reduced to the same metric.

This distinction must not be used as a weapon against inconvenient conclusions. It is easy to call any result that contradicts a preferred belief "scientism." The critique becomes mature only if it preserves epistemic hierarchies. A well-conducted study is not equal to an isolated testimony, and a consensus built through multiple methods is not merely a collective opinion. Meta-rationality does not say that all cones have equal power. It says that their power must be related to the question and to their domain of validation. Outfinitism does not lower standards. It requires the force of the conclusion to be proportional to what has actually been verified.

Metrics make the problem visible because they turn values into administrable objects. A good indicator compresses a complex reality into a useful signal. The number of infections, response time, graduation rate, or energy consumption can reveal failures that intention conceals. But once reward is attached to the indicator, people learn to optimize the number. Goodhart formulated this difficulty in the context of monetary policy, and the observation became an informal law of institutions: when a measure becomes a target, its relation to the value measured degrades. [GOODHART-1975]

A support service that rewards the number of closed tickets can discourage the resolution of difficult cases. A safety program judged by a low number of reported incidents can teach the organization not to report. A research team measured by deliverables can multiply documents and reduce the time available for the investigation that would make them valuable. Each system becomes more efficient within the representation. That very efficiency amplifies the loss between proxy and purpose.

Here hypocrisy acquires an adaptive function. Institutions declare ideals that are simpler and higher than they can implement. “Everyone is equal before the law,” “merit decides,” “science pursues only truth,” “safety is our priority,” or “the user is at the center” are normative compressions. They provide direction and make criticism possible. Real life contains insufficient resources, incompatible roles, interests, attachments, exceptions, and trade-offs. When the distance between ideal and practice cannot be acknowledged, the organization learns to preserve the language of the ideal and move adaptation into opaque zones. Hypocrisy becomes the lubricant of a system that demands impossible purity.

Recognizing this function is not moral justification. Some discrepancies are unavoidable compromises; others are exploitation. The difference depends on power, on who benefits, on externalized costs, and on the possibility of correction. A judge may need discretion to apply a general rule to a particular case. An official who uses the same discretion for favors corrupts the function. A person may express politeness without revealing every inner reaction; a leader who uses the language of care to conceal systematic harm practices extractive hypocrisy. Meta-rationality tries to see both the ideal and the adaptive function. Outfinitism asks which ideal can be implemented under actual conditions, what transition is possible, and which cost can no longer be justified by finitude.

This problem matters because simplistic ideals can produce precisely the behavior they condemn. A culture that demands perfect moral consistency teaches people to hide contradictions. An organization that claims every decision follows procedure creates ceremonial procedures and invisible informal negotiations. A scientific institution that describes research as a purely disinterested process may become less capable of discussing the incentives, prestige, and competition that shape it. In all these cases, absolute language does not eliminate adaptation; it makes adaptation harder to observe.

Rationality can also become an instrument of manipulation without formally violating logic. An actor can select convenient premises, demand from an opponent a level of evidence not required of their own claims, and use uncertainty asymmetrically. An institution can commission analyses until delay serves its interest. A report can delimit the question so that the desired solution appears inevitable. The form of the argument remains rational, but its orientation is strategic. Manipulative rationality does not always lie; it may select truths that lead the interlocutor toward a predetermined conclusion.

AI makes this power cheaper. A model can rapidly generate justifications adapted to the vocabulary of each audience, produce quantities of argument that exceed the capacity to respond, and imitate balance without changing the objective of the institution using

it. A company can offer a personalized explanation of an automated decision, but the explanation may be only a plausible post hoc story. A system can list risks and continue optimizing the same metric. It can simulate empathy while the process offers no real path of appeal. Verbal coherence becomes the decoration of a power that has not changed its frame.

Wisdom is the internal limit of this rationalization. It does not reject the rule, but preserves the connection between the rule and life. It does not reject measurement, but remembers what was compressed. It does not reject science, but separates prediction from permission. It does not reject the ideal, but refuses to turn the impossibility of total compliance into an apparatus of shame and concealment. Meta-rationality provides the mobility needed to see other frames. Outfinitism provides the discipline needed to decide under finite conditions and to declare what remains outside certification. Together they do not diminish reason. They restore it to the status of an extraordinary instrument that becomes stronger when it knows its edge.

CHAPTER 2

IDEAS THAT LEAVE THEIR DOMAIN

Relativity, Quantum Theory, Evolution, and Cybernetics

Some theories acquire a second life after leaving the laboratory, the formula, or the community that produced them. The second life is not a faithful copy. Culture takes an image, a relation, a word, or a tension and reconstructs it according to its own needs. A physical theory can become a metaphor for the experience of time; a biological explanation can become a social ideology; an engineering concept can become a language for family and organization. Transfer is almost inevitable. A culture cannot keep its major instruments of knowledge isolated from the way it imagines the world. The problem is not that transfer exists, but what status it is given.

Einstein's relativity is a privileged example because the distance between the technical theory and its cultural life is easy to observe. In physics, special and general relativity are precise mathematical theories about measurement, frames of reference, spacetime, energy, and gravity. They do not claim that every opinion is equally valid. On the contrary, they seek laws and invariants that allow translation among descriptions made from different frames. Modern culture, however, retained above all the fact that the observer and the frame matter. Literature, art, and philosophy found in relativity an image suited to a world in which social simultaneity was fragmenting, the rhythms of the city were accelerating, traditions were coming into contact, and the single perspective was losing credibility. [WHITWORTH-2001] [GALISON-2003]

This influence should not be judged only by asking whether a novelist understood tensors. Culture was not obliged to become a physics textbook. The transfer could be legitimate as metaphor: human experience depends on position, and two descriptions can differ without reality becoming arbitrary. Yet the popular formula “everything is relative” often lost half the lesson. It preserved the plurality of perspectives and forgot the discipline of transformation among them. Scientific relativity does not eliminate truth; it redefines the conditions under which descriptions can be compared. In meta-rational terms, the fertile lesson is not that every cone has its own isolated truth, but that we must make the frame explicit, find the relations of translation, and search for what remains stable under a change of perspective.

Quantum mechanics produced an even more dramatic migration. Words such as “observer,” “uncertainty,” “entanglement,” “energy,” “frequency,” and “field” entered spirituality, marketing, popular psychology, and alternative medicine. There are genuine

philosophical questions surrounding quantum theory, and the development of quantum technologies shows how concrete the formalism is. But in popular discourse the terms are often detached from mathematics, physical scale, mechanism, and measurement procedure. Prestige remains while constraints disappear. A therapy is called "quantum" without identifying a relevant physical process, a testing protocol, or a prediction that could fail. [BRANG-2025]

We can call this operation **semantic prestige laundering**. A word keeps the authority it acquired in a rigorous domain while losing the obligations that gave it meaning there. Not every metaphorical use is abuse. To say explicitly that entanglement offers a poetic image of interdependence can be fertile, provided one does not claim that a social relation has a quantum mechanism. Abuse begins when a metaphor is presented as an explanation, the explanation as a mechanism, and the mechanism as a justification for intervention. The word crosses clandestinely from analogy to causality.

Darwinism shows another form of escape. Natural selection gives biology a powerful structure for understanding variation, inheritance, and adaptation. Transferred into economics or institutional analysis, the idea of selection can explain why some practices survive without central planning. Cultures, languages, and organizations can be studied through variation and selective pressures. Problems arise when the description of a process is converted into a norm. The fact that a trait survived does not make it morally good. The fact that a company wins in competition does not make the social order that produced the victory just. "Survival of the fittest" becomes a justification of the strongest only by introducing a normative premise that biology does not supply.

The error has a recurring structure. The source domain describes a mechanism. The target domain turns the mechanism into an image of what is natural. The "natural" is then used to legitimize what must be accepted. The transfer hides the movement from is to ought. Meta-rationality makes that movement visible. Outfinitism asks us to declare the limits of the analogy: which relations are preserved, at what scale, and under what conditions the correspondence breaks down.

Cybernetics carried a different family of structures: feedback, self-regulation, control, communication, stability, and recursive loops. [ASHBY-1956] These concepts fertilized biology, psychology, anthropology, organization theory, and family therapy. Here the transfer was sometimes more than metaphor. A biological system and a technical one can indeed have formalizable feedback relations. Yet when the language of control is applied to human beings, dimensions appear that do not exist in a thermostat: interpretation, contestation, dignity, power, anticipation of the rule, and the capacity to change the objective. A manager who treats an organization as a regulatory mechanism may notice

important loops and miss the fact that its members are not merely components, but participants capable of rejecting the meaning imposed on them.

"Information" and "computation" have probably become the most expansive concepts of the digital age. Genes are described as programs, the brain as a processor, the organism as an information system, society as a network, and the universe as a form of computation. Some of these analogies support precise models. Others are vocabularies of orientation. They make coding, transformation, state, and algorithm visible, but can hide the body, energy, history, context, and value. To say that a person "processes information" may be true at one level and almost useless at another. A description that is too general becomes impossible to refute precisely because it no longer discriminates among phenomena.

These examples show that an idea does not travel as a single package. Terms, relations, methods, metaphors, prestige, or normative implications may migrate. Sometimes the structure is preserved while the vocabulary changes. At other times the vocabulary is preserved and the structure is lost. A critic may be right that literature did not adopt the formalism of relativity, and a cultural historian may be right that relativity changed the modern imagination. They are discussing different kinds of influence. To judge a transfer, we must ask what exactly was transferred and what claim is assigned to it in the new domain.

Responsible transfer	**Abusive transfer**
Declares whether it is an analogy, model, or mechanism and does not change status clandestinely.	Uses ambiguity to cross from metaphor to causality.
States which relations are preserved and where correspondence breaks down.	Preserves only the words or prestige of the source domain.
Accepts tests and counterexamples proportional to the claim.	Becomes harder to refute as the claim grows.
Does not turn description into a norm without explicit axiological premises.	Presents what seems "natural" as what must be accepted.

We do not need a police force for metaphors. Culture would become sterile if every analogy required a mechanistic demonstration. Poetry, philosophy, and scientific invention use comparisons to open spaces of thought. The requirement is more modest

and more important: do not use the freedom of metaphor to borrow the authority of mechanism. The greater the consequences - medical treatment, public policy, evaluation of a person - the stronger the bridge between domains must be.

Memes, Attractors, and the Ecology of Reconstruction

The metaphor of the meme offered a memorable image for the cultural life of ideas. Contents would compete for attention, memory, and transmission, almost like replicators in an ecosystem. [DAWKINS-1976] The image is fertile because it shifts the question from the author's intention to population dynamics. An idea may survive not because it is true, but because it is easy to remember, confirms identity, produces emotion, benefits from prestige, or attaches itself to an institution. It can combine with other ideas into an ensemble whose parts protect one another. It can offer justifications for behaviors that secure its own reproduction.

Taking the metaphor literally creates problems, however. Ideas are not organisms with metabolism and intentions of their own. Human beings do not passively copy identical representations. They reconstruct what they hear according to memory, interest, language, and expectation. Research on cultural attractors shows that the stability of a form can arise not only through faithful replication, but through convergent transformations. Different variants are reconstructed in similar directions because certain forms are easier to understand, remember, or legitimize in a given environment. [SPERBER-1996] [BUSKELL-2017] [CLAIDIERE-SPERBER-2007]

This observation explains why sophisticated theories enter culture in compact forms. Relativity becomes “truth depends on perspective.” Evolution becomes “the strong win.” Quantum theory becomes “the observer creates reality.” Cybernetics becomes “everything is feedback.” Neuroscience becomes “the brain compels us.” These formulas are attractors. They compress a tension into a sentence that is easy to use. They are not always completely false; it is precisely the mixture of truth, exaggeration, and usefulness that makes them resilient.

We can speak of a memetic ecology if we preserve the metaphorical character of the term. Ideas activate and inhibit one another. A belief about merit can support a belief about markets; a belief about identity can decide which sources are trustworthy; a story about progress can make present costs appear as unavoidable investments. Institutions provide habitats. The university selects through curricula, journals, and reputation. The platform selects through recommendation and rewards of attention. The state selects through legal categories. The company selects through indicators and promotions. An

idea embedded in software or procedure no longer depends on the explicit conviction of every participant. It continues to organize behavior through infrastructure.

For this reason, the "cone of meaning" should not be located exclusively inside the head. A perspective is distributed among vocabulary, habit, instrument, role, and institution. Legal rationality lives in the mind of the lawyer, but also in codes, archives, precedents, deadlines, and avenues of appeal. Science lives in the researcher, but also in instruments, protocols, databases, peer review, and reporting standards. Capitalism or nationalism are not merely doctrines that people accept; they are practices, rights, borders, contracts, symbols, and incentives. A cultural cone becomes stable when the environment makes it easy to carry even for those who do not know its theory.

This helps us understand historical epochs without reifying them. An epoch does not have one mind. Dominant cones coexist with residual traditions, subcultures, and counter-frames. Yet institutions can make some questions easy to formulate and others almost impossible to hear. A researcher may find it natural to look for mechanism, an administrator for an indicator, a lawyer for competence, and an activist for the distribution of power. Each question can reveal something real. The ecology becomes impoverished when one conceptual species occupies every niche.

The virality of an idea is not evidence of truth. It is evidence that the idea fit an environment of transmission. Some true ideas are hard to communicate; some false ideas satisfy powerful psychological and institutional needs. But it would be equally mistaken to treat virality as superficial. An idea that reorganizes institutions produces social reality even if its description of the world is false. A pseudoscientific racial theory can be empirically refuted and still structure laws, distributions of resources, and suffering. Cultural effect and epistemic truth are two different axes.

The ecology of ideas does not eliminate responsibility. The language of memes can suggest that no one decides anything because representations simply "spread." In reality, actors with power select what is amplified, funded, taught, and implemented. Platforms choose recommendation functions. Institutions choose standards. Leaders choose metaphors. Individuals choose when to retransmit and when to correct. Systemic dynamics explain why outcomes exceed individual intention, but they do not absolve actors who can intervene.

The same principle clarifies hypocrisy without absolving it. A culture of ideals that cannot be implemented selects behavior for show, but the moral position depends on power. The subordinate who simulates enthusiasm in a coercive environment and the leader who simulates transparency to preserve control respond to different pressures and produce different costs. Ecology describes the mechanism; axiology preserves the asymmetry.

The language model enters this ecology as a machine of reconstruction. It was trained on traces of many cones and can produce new variants that fit cultural attractors. It can make an idea more accessible, translate between technical and ordinary vocabulary, and discover fertile analogies. It can also accelerate semantic degradation. A simple and persuasive formulation can be produced in millions of variants, adapted to every audience, and separated from the source that would have constrained it. AI does not invent memetic ecology, but it lowers the cost of reproduction and mutation.

A requirement for generative culture follows: literacy no longer means only evaluating the proposition. We must notice which cone it carries, which prestige it borrows, what change of status it performs, and which institution amplifies it. An idea can be true and used manipulatively. It can be false and symbolically fertile. It can be a good metaphor and a poor mechanism. Meta-rationality separates these functions. Outfinitism requires the force of use not to exceed the bridge of validation. Axiology asks what happens to human beings after the idea becomes infrastructure.

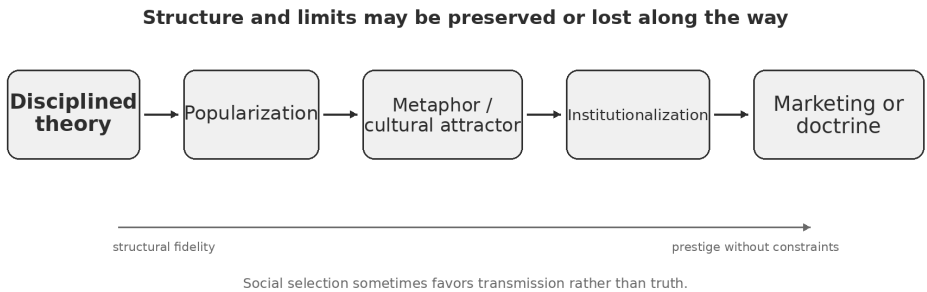

Diagram 2. The cultural life of a theory

CHAPTER 3

CONES OF MEANING AND THE EPOCHS THAT VIEW THE WORLD THROUGH THEM

Geometry as a Disciplined Metaphor

The metaphor of the cone emerged from a modest technical observation. In many systems of semantic representation, words, sentences, or documents are associated with vectors in a high-dimensional space. Cosine similarity compares the direction of these vectors, rather than only their raw distance. If we choose a reference direction and consider representations close when their angle to it remains below a threshold, the resulting region has a conical form. The points inside it are not identical, but they are oriented similarly enough to be treated as neighbors relative to that reference.

This property does not prove that meaning is literally a cone or that inside an LLM there are clean geometric objects corresponding to philosophies. Contextual representations change from one layer to another, the same word occupies different positions in different contexts, and the global geometry may be anisotropic and difficult to interpret. [ETHAYARAJH-2019] In a large model, behavior is produced by successive transformations, attention, distributed activations, and conditional probabilities. The cone is a compression constructed by the observer, not a photograph of the internal mechanism.

Even so, the image preserves something valuable. A semantic orientation does not need an exhaustive definition in order to produce coherent consequences. If a prompt draws together ideas about finitude, the plurality of frames, epistemic caution, verification, wisdom, and the possibility of extending limits, the model can continue in a direction that connects them. Many different sentences enter the same field of compatibility. Some will emphasize philosophy of science, others psychology, mathematics, governance, or AI. The cone does not determine one sentence; it makes a family of continuations more probable.

The idea also applies to human beings. A person looking through an economic grid notices incentives, costs, scarcity, and exchange. A legally oriented person notices competence, rule, evidence, and procedure. A therapeutic orientation makes suffering, relationship, and psychological defense visible. An engineering orientation searches for

function, mechanism, reliability, and failure. A religious orientation may select meaning, sin, hope, ritual, and transcendence. These cones are not merely opinions about the same list of facts. They help construct the list. They make some entities appear as relevant objects and leave others in the background.

In a labor conflict, the legal cone may ask whether procedure was followed. The economic cone may ask what cost the decision creates and what signal it sends. The psychological cone may notice fear and status. The moral cone may ask who is being treated as a means. The organizational cone may see the effect on trust. None of these questions is automatically sufficient. A decision can be legal and destructive, efficient and unjust, empathetic and impossible to implement. Meta-rationality begins when we can distinguish these orientations and discuss why one should receive priority in a case without declaring the others unreal.

The fragility of the cone comes from the plurality of reference directions. In a high-dimensional space, the same representation can be close to several directions. A word such as "freedom" can enter a legal, moral, economic, psychological, national, or spiritual cone. Context decides which neighborhoods become active. A small change of role can reorganize meaning: "freedom" in a judicial ruling does not function as it does in an advertisement or a confession. Neither people nor models necessarily carry a single stable orientation. A researcher can be rigorous in the laboratory and dogmatic in politics; a lawyer can think procedurally at work and relationally at home. Cones overlap, are activated situationally, and may conflict within the same person.

The cone therefore describes a tendency of selection, not a hidden essence. Sometimes the network, the field, the role, the practice, or the narrative is the better image. A meta-rational concept must allow itself to be replaced; if it explains everything, it has lost its power of discrimination.

With this caution, the cone offers a more accurate image than the lens in two respects. The lens suggests that there is a complete scene that the instrument merely distorts or brings into focus. The cone suggests the selection of a region from a broader space of possibilities and relations. In addition, a lens appears to be a singular object, whereas cones can be multiple, overlapping, and differently oriented. A culture can simultaneously carry the scientific cone of evidence, the legal cone of rights, the economic cone of growth, and the religious cone of meaning. Problems arise not only from the error of each, but from the implicit hierarchy among them.

The cone also suggests that meaning has porous edges. Some expressions lie near the axis and are easily recognized as representative. Others are peripheral, where membership depends on threshold and context. A doctrine has canonical texts and marginal interpretations. A culture has central practices and hybrids. A model may respond

reliably to typical examples and behave unpredictably in boundary cases. Alignment is often most difficult precisely at the edge: when rules conflict, intention is ambiguous, risk is indirect, or the context does not resemble the examples used in training.

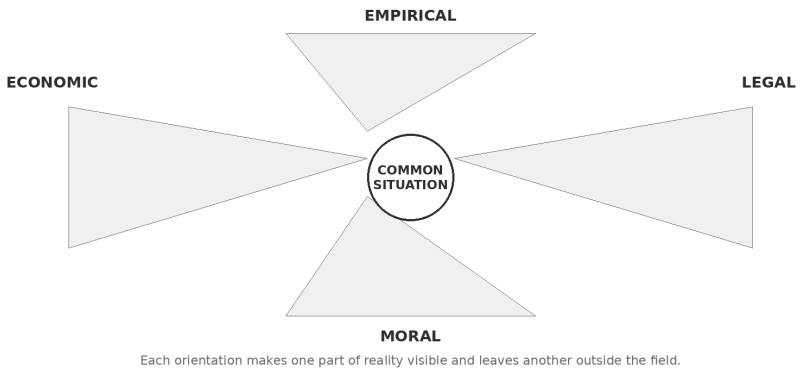

Diagram 3. Several cones over the same situation

The metaphor thus becomes a proposal for literacy. Understanding a text does not mean only paraphrasing its propositions. It means recognizing which orientation it asks the reader to adopt, which relations it makes natural, which alternatives become difficult to think, and which authority it borrows. An advertisement may carry the cone of individual choice while concealing the infrastructure that restricts options. A report may carry the cone of measurement and make subjective experience look like noise. An ideology may carry the cone of conflict and interpret every objection as a symptom of the adversary. In the generative age, this literacy becomes even more important because a model can rapidly construct coherent texts from within any requested orientation.

Epistemes, Cultures, and Institutions: When the Cone Becomes a World

Philosophy and history of science have offered several concepts close to what we call a cone of meaning. Thomas Kuhn described paradigms through which scientific communities share exemplary problems, methods, and criteria of success. [KUHN-1962] Ludwik Fleck spoke of thought styles and collectives that make certain facts possible as objects of knowledge. [FLECK-1935] Ian Hacking showed that styles of reasoning provide not only methods, but also types of propositions, objects, and explanations that become intelligible. [HACKING-1992] Michel Foucault used the notion of episteme to describe the historical conditions that organize the possibility of discourse in an epoch. [FOUCAULT-1966] Karin Knorr Cetina studied different epistemic cultures in which

sciences construct knowledge through distinct practices and objects. [KNORR-CETINA-1999]

These traditions are not identical and should not be compressed into a single concept. Together, however, they support an important idea: modes of knowing are not merely collections of claims. They are practices that determine what can appear as a problem, object, evidence, and explanation. A particle-physics laboratory, a clinic, a courtroom, and a historical archive do not merely use different data. They construct different relations among observer, instrument, subject, and validation. Requiring all of them to imitate the same model of objectivity might produce not unity, but the loss of the object.

Historical epochs can be viewed as ecologies in which some such cones acquire a dominant position. Medieval Europe was not reducible to religion, and modernity is not reducible to science. Yet the change in sources of legitimacy is real. In a theological order, questions of ultimate meaning and order could structure law, politics, and nature. Modernity shifted the center toward observation, calculation, the individual, progress, the market, the state, and procedure. The industrial age made mechanism and production general metaphors. The information age has made the network, code, and optimization nearly ubiquitous.

To say that an epoch "sees" through a cone is an abbreviation. It contains conflicts, minorities, and transitions. Old grids do not disappear when new ones arise. Religion coexists with scientism, tradition with individualism, romanticism with algorithmic administration. A person may inhabit several epochs at once: use modern medicine, practice traditional rituals, think economically at work, and morally at home. This overlap produces tensions and hypocrisies. Society declares legal universalism but preserves particular loyalties. It celebrates merit but functions through networks. It invokes individual autonomy but optimizes behavior through persuasive infrastructures.

Cones become durable when they are embedded in institutions. An idea of merit gains force through examinations, résumés, rankings, and selection procedures. An idea of risk becomes reality through policies, scores, and thresholds. An idea of health is materialized in diagnostic codes, protocols, and reimbursement. An idea of productivity is installed in dashboards, quarterly objectives, and monitoring software. Participants do not have to believe explicitly in the philosophy of the system. It is enough that they adapt to its rewards.

This process explains how a grid can produce its own confirmation. If a school defines intelligence by what a test measures, the curriculum will be reorganized to raise the score, and the score will later appear to be an even better description of intelligence. If a platform defines relevance through engagement, producers will adapt, and engaging

content will occupy more of reality. If the market defines value through price, activities that are difficult to monetize will receive fewer resources, and the lack of investment will appear to prove their lower value. The model does not merely see the world; it participates in configuring it.

This reflexivity is especially powerful in the social domain. A planet does not change its orbit because it has read the astronomer's theory. A person can change behavior after learning how they have been classified. An institution can exploit the rule by which it is measured. A prediction can become self-fulfilling or fail because actors reacted to it. The limit of the model then lies in the loop between representation and world, not only in the initial data.

This is why we need an ecology of models. It is not enough to keep multiple perspectives in a tolerant museum. They must be able to correct, translate, and limit one another. The economic model can reveal incentives that moral discourse ignores. The moral model can reveal costs concealed by the economic aggregate. Law can create procedures of contestation; lived experience can show that procedure no longer serves justice. Science can correct factual beliefs; politics and ethics must decide what authority the result receives.

Plurality does not guarantee the health of the ecology. Several theories may share the same assumptions or depend on the same infrastructure. Multiple AI models may have similar data, architectures, and incentives. An institution may invite different perspectives only to create the appearance of consultation. Diversity becomes real when there are differences of method, interest, access to power, and the possibility that an objection can change the decision.

Here meta-rationality acquires a political meaning. It is not merely an individual's ability to “see several sides.” The question is who has the right to define the sides, who can challenge the dominant cone, and what happens when translation produces losses. An expert can understand multiple frames and still select only those compatible with the institution's interest. An organization can speak sophisticatedly about complexity while preserving the same distribution of power. Wisdom is not measured by the number of perspectives listed, but by the willingness to let a relevant perspective modify action.

Outfinitism adds the discipline of stopping and consequence. Society cannot consult every cone indefinitely: the physician, the court, and the administrator must decide. Yet the decision can carry the memory of its limits by remaining monitored, contestable, and reversible. Finitude does not justify eliminating inconvenient voices; it requires more honest prioritization.

Cones of meaning are useful not as labels for cultures, but as a way of making the relation between orientation and blindness visible. Every cone permits seeing and produces an edge. Wisdom begins when people and institutions can make the cone itself an object of vision.

CHAPTER 4

META-RATIONALITY: THE WISDOM THAT SEES THE FRAME

Purpose, Scale, Time, and Power

Meta-rationality can sound like a step above rationality, as though there were a privileged observer who sees every system without belonging to any of them. That interpretation would contradict the very idea the term is meant to express. No person, culture, or machine stands outside language, history, interests, and limits. Meta-rationality is not a view from nowhere. It is the situated capacity to include, for a time, the current frame within the field of criticism. One frame examines another and can in turn be examined. We gain distance, not omniscience.

In its simplest form, rationality asks whether a conclusion follows from premises and whether an action serves the stated objective. Meta-rationality adds questions about the premises, the objective, and the domain: is this what we are actually trying to protect? Have we formulated the problem at the right scale? What does the model make visible, and what does it remove? Who chose the criterion of success? Under what conditions should the frame be changed? What other perspective would alter the decision? These questions do not replace calculation. They decide when calculation is being applied to the right problem.

The companion volume defined meta-rationality as the capacity to reason rigorously while examining the frame that makes the act of reasoning possible. [ALBOAIE-2026] The definition contains two parts that must remain together. Without rigor, the "meta" becomes a refuge for impressions that refuse verification. Without examination of the frame, rigor can optimize an error of formulation. Meta-rationality is the mature relation between these two requirements: enough commitment to act and enough distance to correct.

A difficult medical decision shows why this relation can be called wisdom. Clinical literature can estimate probabilities, and the physician can explain effects. But the choice of a treatment that modestly prolongs life at the price of pain and lost autonomy is not decided by statistics without the person. The model of survival, the model of quality of life, family responsibilities, uncertainty in the evidence, the image of dignity, and tolerance for risk are all relevant. A rational process must respect the facts. A

meta-rational process must notice why different facts matter for different purposes and where no optimization eliminates judgment.

Research on reflective judgment, postformal thinking, and wisdom studies related capacities: recognition of uncertainty, evaluation of sources, integration of perspectives, sensitivity to context, and the connection of knowledge to values and consequences. [KING-KITCHENER-1994] [KING-KITCHENER-2004] [BALTES-STAUDINGER-2000] These traditions do not confirm the existence of a universal psychological stage called meta-rationality. The term should be used as a philosophical synthesis and a family of practices, not as a title of superiority. A mathematician can be meta-rational about formal models and rigid in personal relationships. A leader can understand geopolitical complexity and remain unable to revise an image of the self. The capacity is situated, variable, and exposed to regression.

The first meta-rational movement is epistemic: distinguishing what we know, what we infer, and what we assume in order to act. Uncertainty is not a single state. We may have insufficient data, imprecise measurement, disagreement among models, instability of the environment, ignorance of the relevant values, or uncertainty about our own competence. Each calls for a different response. Missing data may call for research; conflict among values may call for deliberation; instability may call for a reversible action; high stakes may call for escalation to a more appropriate authority. The generic formula "we are not sure" conceals these differences.

The second movement concerns purpose. Instrumental rationality often receives the objective as an input. Meta-rationality periodically returns to the question the system has forgotten: what are we trying to protect, and is the indicator still connected to that thing? A university may begin by measuring publications in order to observe research activity and end by organizing research to maximize publications. A company may measure customer satisfaction and learn to obtain high scores by pressuring respondents. A safety policy may begin as protection and become a mechanism of control. Meta-rationality does not assume that objectives must be changed constantly. It creates moments in which the objective ceases to be invisible.

The third movement concerns scale. An action that is reasonable for an individual can be destructive when repeated by millions. Driving to work may be an individually rational choice and can collectively contribute to congestion. A company may believe it cannot afford to slow a dangerous race, even though every company would prefer a world in which the race were constrained. A household buys the cheapest product and, together with millions of similar decisions, supports labor conditions it would not accept as a public rule. Local rationality does not automatically compose into collective rationality.

The reverse is also true. A policy can improve the average while imposing severe losses on a small group. A treatment reduces population risk but harms a subgroup whose characteristics were poorly represented. A city becomes more prosperous in the aggregate and expels those who contributed to its attractiveness. The average can be correct and morally insufficient. Scale is not a technical detail added after the result; it changes the meaning of the result.

Time can likewise reverse judgment. A strategy increases quarterly profit and consumes the capacity to innovate. An administrative exception solves an emergency and becomes the permanent form of power. A technology offers immediate efficiency and creates institutional dependence. Meta-rationality asks not only whether the solution works, but for how long, with what learning effects, and which options it closes. A decision can be correct now and still need an expiration condition.

Power is the dimension that purely cognitive discourse can miss. A frame is not only a representation; it can be a distribution of authority. Whoever defines the problem decides which suffering becomes relevant. Whoever controls the indicator can declare success. Whoever has access to the exception lives under a different system from the person forced to follow the rule literally. Epistemic trust becomes command when it is attached to an institution. The less opportunity a person has to appeal, the more explicit, justified, and contestable the frame that classifies them must be.

Meta-rationality is therefore also an ethics of power. An expert need not abandon expertise in order to recognize that applying it entails a relation. A physician knows more about the mechanism of treatment; the patient knows something else about what may be lost. An engineer can estimate reliability; the affected community can identify values that were not included in the objective. A court has the authority to decide, but its legitimacy depends on procedure, justification, and appeal. Not every voice has equal competence about every claim, but technical competence does not automatically confer total moral authority.

The position can be summarized without a mechanical checklist. Meta-rationality makes the frame visible, separates types of disagreement, seeks a genuine alternative, compares consequences, and decides at the level of resolution permitted by the stakes. After the decision, it preserves the possibility of correction. It does not require every problem to become a philosophical seminar. Sometimes the frame is stable, the procedure has been validated, and further reflection would reduce safety. A pilot does not reconstruct the philosophy of aviation during an emergency. Maturity includes knowing when a change of frame adds no value.

Pluralism Without Relativism, Decision Without Omniscience

Recognizing frames immediately raises an objection: if every perspective has its own assumptions, do we not reach the conclusion that all are equivalent? That would be a lazy form of relativism and a betrayal of meta-rationality. Frames can be compared. Some make predictions that fail. Some contradict robust evidence. Some are internally incoherent. Some work at one scale and become wrong at another. Some are effective and morally unacceptable. The fact that evaluation also has a frame does not erase the difference between a testable argument and a claim that changes its meaning to avoid refutation.

Mature pluralism says that several models may be necessary for complex objects. It does not say that every model is good. Physics uses different descriptions at different scales and links them through mathematical relations. Biology integrates molecular mechanisms, development, behavior, and evolution without automatically reducing every question to a single level. The social sciences need quantitative data, history, institutions, and interpretation of meaning. [CARTWRIGHT-1983] [GIERE-2006] [MITCHELL-2009] The criteria are not identical in every domain, but each domain can require discipline, transparency, and contact with its object.

Meta-rationality does not suspend judgment; it makes judgment more articulate. Instead of declaring "this is the final truth," it can say: this model is preferred here because it explains better, predicts, enables intervention, reduces certain errors, and has acceptable consequences; we will not extend it beyond these conditions without new evidence. Such a statement is not weaker. It is more informative. It separates justified confidence from the psychological need for certainty.

Decision under uncertainty is where meta-rationality approaches wisdom. A person can wait for more information until the moment for action has passed. They can generate alternatives indefinitely and call paralysis "complexity." Meta-rationality does not wait for omniscience. It seeks a decision whose force is proportional to the evidence, whose reversibility fits the stakes, and whose cost of error has been considered. Sometimes the wisest option is a small action, an experiment, or a stage that preserves alternatives. At other times a short deadline requires a decisive choice. Knowledge of limits does not eliminate courage; it frees courage from the false loan of certainty.

Emotion has an ambivalent role in this process. Simplistic rationalism treats it as noise added to a mind that would function better as a computer. Yet emotion can mark value, danger, loss, and relevance before analysis is complete. [DAMASIO-1994] Fear can signal risk and distort probability. Shame can support social conscience and impose conformity. Joy can indicate alignment with a value and reward self-deception.

Meta-rationality does not enthrone feeling. It treats feeling as information to be interpreted, neither eliminated nor absolutized.

Identity is equally important. A belief tied to belonging is not revised merely by presenting a statistic. Abandoning it may threaten relationships, status, and the coherence of the self. This fact explains resistance, but does not prove the truth of the belief. A meta-rational practice seeks forms of correction that do not unnecessarily destroy the person. At the same time, it does not turn protection of identity into a veto over reality. Wisdom can preserve dignity while separating experience from factual interpretation.

Hypocrisy enters here not as a marginal exception, but as a symptom of conflict between general rules and particular lives. Ideals coordinate, yet no actor can apply them without interpretation, and interpretation inevitably distributes costs and exceptions.

The problem is not exhausted by detecting inconsistency. We must ask who benefits, who bears the cost, how much power the actor has, and whether the adaptation preserves the function of the ideal or uses it as a screen. The hypocrisy of the constrained and the hypocrisy of those who control the rules are not symmetrical. Causal understanding is not absolution.

There is also a counterfeit of meta-rationality. It uses the vocabulary of frames to avoid every commitment. Every fact becomes "a perspective," every criticism is relativized, and self-interest presents itself as sophisticated integration. A person can explain why everyone is partly right precisely in order to avoid saying where a limit must be imposed. Complexity becomes status. In institutional settings, consultation with several perspectives can become a ritual that legitimizes a decision already taken.

The defense is not an algorithm, but a test of friction. A perspective is meta-rational if it can formulate the strongest counter-frame, say what evidence would change the conclusion, and allow a relevant objection to modify action. If plurality produces no possibility of change, it is decoration. If uncertainty is invoked only when it threatens self-interest, we have strategy, not wisdom.

Rationality, meta-rationality, and outfinitism	**Characteristic question**
Rationality	How do we solve the problem correctly within the given frame?
Meta-rationality	Which frame are we using, what does it exclude, and when should it be revised?

Rationality, meta-rationality, and outfinitism	Characteristic question
Outfinitism	What can we support and do under actual limits, and how can the frontier be pushed responsibly?

The table does not describe stages that cancel one another. Meta-rationality needs rationality, and outfinitism needs both. A good decision does not remain permanently above the problem. It enters a frame, calculates, returns to consequences, and knows when to reopen the question. The movement resembles breathing more than climbing a ladder. Inhalation expands the field; exhalation closes provisionally and makes action possible.

Meta-rationality can be called wisdom precisely because it refuses two illusions: the illusion that we need certainty in order to act, and the illusion that recognizing uncertainty absolves us from choosing. It does not promise reconciliation among all values. Sometimes there are tragic conflicts, irreducible losses, and frames that cannot be composed without residue. In such situations, its role is to prevent the loss from being hidden under the language of technical necessity. A decision may remain necessary and still acknowledge what it sacrifices.

Wisdom is not a larger database. It is the right relation to what we know, what we do not know, and the power a conclusion receives. A person can have the correct answer within the wrong frame and produce a sophisticated error. They can have an incomplete model and, through modesty, monitoring, and the possibility of correction, act better. Meta-rationality does not guarantee the outcome. It creates better conditions for error to become visible before it becomes destiny.

CHAPTER 5

OUTFINITISM: THE ART OF OPEN LIMITS

Actual Finitude and the Open Horizon

Meta-rationality makes the frame visible. Outfinitism asks what can be done after we have seen it. The first frees us, at least provisionally, from the impression that one perspective exhausts reality. The second prevents us from turning that freedom into an imaginary journey through every possible perspective. Every comparison consumes time. Every model requires data, language, and instruments. Every verification has a resolution, duration, and cost. Every institution can sustain only a limited number of procedures, exceptions, and forms of appeal. Outfinitism begins from this concrete finitude, but refuses to confuse it with a metaphysical closure of the world.

The book to which this volume is added called this position "the open finite." [ALBOAIE-2026] The expression is meant to hold together two truths that modern culture often separates. The first is that every effective achievement is finite. A proof is written in a language and checked through determinate operations. An experiment measures a limited domain. A language model has an architecture, a context, a memory, and access to particular tools. A public policy is implemented by concrete people, over a period of time, with a budget and an administrative capacity. The second truth is that these achievements do not necessarily close the domain. We can continue an investigation, revise a theory, extend a simulation, modify an institution, or invent a new language without assuming that all future stages already exist as a complete inventory.

This position is not a rejection of the mathematics of infinity, nor a return to the idea that only small or immediately constructible objects are legitimate. Modern mathematics has produced rigorous theories of infinity, and their use should not be challenged by vague intuitions about the physical world. Confusion begins when the prestige of a formal abstraction is transferred clandestinely to cultural promises of growth without cost, total control, complete prediction, absolute security, or intelligence without limits. The fact that a mathematical system can describe an infinite totality does not prove that an ecosystem can support endless economic expansion. The fact that a process can be improved repeatedly does not mean that perfection is an available product. The fact that a model becomes larger does not grant it unlimited authority.

Outfinitism is less interested in banning a word than in demanding a translation. When someone promises the "unlimited," the question becomes: unlimited relative to which operation, for which agent, over what interval, with what resources, and in what sense? Some limits are physical, others epistemic, economic, legal, moral, or institutional. Some are relatively stable; others move when new instruments appear. Some must be accepted, others compensated for, negotiated, or overcome. Naming a limit does not mean sanctifying it. It means turning an unclear edge into an object of judgment.

Some limits function as walls. The speed of light, biological mortality under present conditions, or the inability of an institution to process an unlimited number of cases simultaneously can impose hard constraints relative to a task. Other limits are thresholds. A team can manually verify ten artifacts carefully and a thousand only superficially. A model can work well until the data distribution changes. A treatment can be useful below a dose and toxic above it. Other limits are moving horizons: what could not be measured yesterday becomes accessible through a new instrument; what could not be computed within a budget becomes practical after an algorithmic improvement. There are also normative limits, which express not inability but refusal: even if we can completely surveil a population, it does not follow that we should.

A mature culture does not reduce all these limits to the same proposition. "We cannot" can mean logical impossibility, temporary technological absence, excessive cost, lack of legitimacy, insufficient evidence, or simply lack of political will. Confusing them often serves power. A company can present an economic decision as a technical necessity. A state can call an institutional choice inevitable. A researcher can turn the limits of an experiment into limits of reality. Outfinitism requires the type of limit to be explained, precisely so that we can see whether it should be respected, attacked, or redistributed.

In this sense, outfinitism is an art of the frontier, not a metaphysics of capitulation. It requires every important claim to carry the memory of the conditions under which it was obtained. A conclusion about the effectiveness of a treatment should state which populations were studied, which indicator was used, and which effects remained poorly observed. An economic prediction should indicate the scenario and interval. An AI system evaluated on benchmarks does not become "safe in general"; it becomes a system that passed particular tests under particular conditions. The limit is not an embarrassing footnote, but part of the meaning of the claim.

This discipline responds to a change produced by AI. Generation has become much cheaper than verification. In an hour, a model can propose more hypotheses, proofs, programs, and interpretations than a human team can inspect seriously. In mathematics and science, an artifact can be formally correct and still express the wrong problem, use an inadequate representation, or fail to connect with the empirical world. The volume on

outfinitist mathematics formulated precisely this difference between derivation and operational capability: a proposition may be proved within a system, but using it in a computer or experiment requires additional representations, procedures, budgets, and checks. [ALBOAIE-2026-MATH]

The idea can be generalized without turning all culture into software. A claim does not become mature merely because it sounds coherent. We need to know which questions it answers, what kind of evidence supports it, how far it can be extended, and what remains open. This “open frontier” is knowledge, not failure. Saying that a study does not permit extrapolation to all populations is positive information. Saying that a model cannot distinguish between two causal explanations is more useful than making an arbitrary choice. Saying that a decision should be reviewed after six months may be more rational than presenting it as a permanent solution.

Outfinitism does not require equal caution in every situation. A claim about an aesthetic preference does not need the same validation effort as a medical recommendation. A reversible suggestion can be offered on the basis of an approximation. A decision that can deprive liberty or cause irreparable harm requires another level of verification and another distribution of authority. Finitude does not flatten the world; it demands proportionality. Deliberative resources are limited and should be directed toward the places where error costs most.

This brings outfinitism close to Herbert Simon's bounded rationality, but the emphasis is different. [SIMON-1955] Bounded rationality describes agents that cannot optimize perfectly and therefore use rules, approximations, and satisfactory thresholds. Outfinitism adds an explicit memory of the frontier: the agent not only decides under constraints, but should know which part of the problem could not be represented, what may change, and what authority the decision has. A good-enough choice should not present itself as the absolute optimum. It may be preferable because it is reversible, monitorable, and compatible with later learning.

Here lies the difference between simplicity and simplism. Simplicity is conscious compression. It leaves details aside to make action possible, but preserves the memory of what it removed. Simplism forgets the loss and turns the instrument into a world. A simple medical protocol can save lives when time is short, provided there are criteria for exceptions and escalation. The same protocol becomes simplistic when unusual cases are treated as defects of the patient. A simple legal rule can reduce arbitrariness; it becomes simplistic when literal application destroys the purpose of the law. An AI model can use an effective heuristic; it becomes dangerous when the score is treated as the person's identity.

Outfinitism is therefore not a cult of limits. Some limits are unjust and must be overcome. Lack of access to education, political exclusion, or artificial barriers imposed on a group do not acquire legitimacy merely because they are real. In other cases, the attempt to eliminate every limit produces domination. An organization that demands constant availability treats the employee's body as an optimization problem. A platform that pursues engagement without a threshold treats human attention as an apparently unlimited resource. A society that refuses to accept vulnerability can produce systems of surveillance more destructive than the risk they are meant to fight.

The art of open limits consists in distinguishing. A limit can be accepted, moved, compensated for, shared, transformed, or refused. No single rule decides. Outfinitism offers an orientation: make the limit visible, describe the relation in which it appears, do not confuse the model with the world, and keep open the possibility of a better form. In this formulation, it is closer to practical wisdom than to a doctrine about the size of the universe.

Yin and Yang: The Mobility of Frames and the Discipline of the Frontier

Meta-rationality and outfinitism can be viewed as two complementary movements. The yin-yang metaphor is useful only if we do not load it with exotic claims. It expresses a mutually corrective tension. Meta-rationality opens: it searches for other frames, translates among perspectives, makes assumptions visible, and resists premature closure. Outfinitism closes provisionally: it remembers that analysis has a budget, that a decision must be made, and that every claim must be tied to actual coverage. The first without the second risks becoming mobility without commitment. The second without the first risks becoming discipline in the service of the wrong frame.

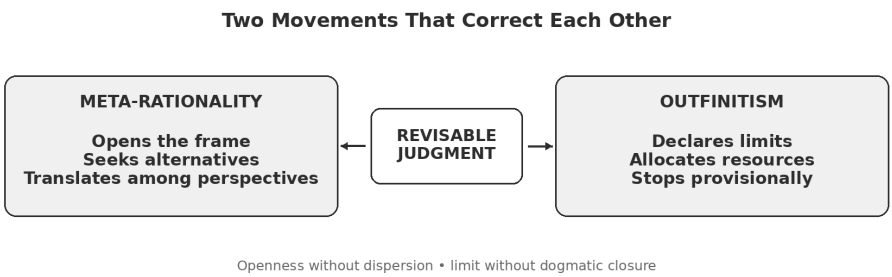

Diagram 4. The two movements of judgment

In a concrete problem, these movements alternate. A research team defines and uses a model. Results appear that do not fit. Meta-rationality asks whether the anomaly is noise, a measurement limit, or a sign of an inadequate frame. Alternatives are proposed. Outfinitism then asks which alternative can be tested with the available resources, what result would count, and when the experiment should stop. After the test, the frame can be preserved, narrowed, or replaced. Neither opening nor closing is final.

The same alternation appears in politics. A general rule is necessary for predictability. Concrete cases show that the rule sometimes produces injustice. Meta-rationality makes visible the purpose of the law, the experience of the person, and other interpretations. Outfinitism remembers that the institution cannot rejudge the entire social order from the beginning in every case. Finite criteria are needed for exception, appeal, and precedent. Wisdom is neither blind application nor permanent arbitrariness; it is the design of a relation between rule and correction.

Social hypocrisy can be analyzed through the same pair. Ideals compress a moral direction and coordinate expectations, while real life produces compromises, roles, and differences between what can be said publicly and what can be done. Purist criticism treats every deviation as evidence that the ideal is false; cynical realism concludes that ideals are only instruments of power. Meta-rationality searches for the function and the hidden tension. Outfinitism asks what degree of compliance is possible, what the discrepancy costs, and what concrete step can reduce the need for appearance.

The formula “preserve the function, withdraw the deception” remains only a rule of orientation. It asks us to distinguish conventions that protect cooperation from forms of dissimulation that protect privilege. Outfinitism brings the ideal into relation with actual capacity, while meta-rationality checks whether the function being invoked is, in fact, the very problem that should be changed.

In science, the pair corrects two opposite excesses. Scientism extends scientific methods to the claim that only what they can currently measure is real or valuable. Anti-science answers with generalized suspicion and treats the limits of a theory as proof that every alternative intuition is equally good. Meta-rationality recognizes the plurality of models and the role of values in choosing problems. Outfinitism preserves the epistemic hierarchies created through experiment, replication, and verification, but ties them to their domains. Not everything is measurable in the same way, yet a causal claim about a treatment does not become true through inner experience.

This complementarity can also be formulated as a relation between epistemology and axiology. Epistemology asks what is justified. Axiology asks what matters and what is worth pursuing. Meta-rationality shows that the questions are connected: the choice of evidence and purpose takes place within a frame. Outfinitism shows that the connection

must be implemented through finite choices. We cannot evaluate all values and all consequences at once, but we can declare what we prioritized, what we omitted, and who bears the risk.

This is decisive for AI. A generative system does not merely produce claims; it distributes attention. Through the order of its answer, the examples chosen, and the tone of certainty, it makes some possibilities more visible than others. A model trained to be helpful may fill every space with answers. An outfinitist filter would introduce a stopping question: is another generation, a longer explanation, or an automated decision necessary? A meta-rational filter would introduce a perspective question: what frame has been assumed, and which relevant alternative is missing? Together they do not provide the machine's final morality, but they may change the way it administers the power to generate.

The danger is that the two concepts become a verbal ritual. A system can declare that "all models have limits," list three perspectives, and continue exactly the same optimization. An institution can add a section on uncertainty without allowing the results to change the decision. A limit is real only if it changes authority, resource allocation, reversibility, or the scope of the conclusion. An alternative perspective is real only if it can produce an observable difference.

Outfinitism must therefore be severe with its own claims. It is not yet a discipline stabilized by a broad community, experiments, and institutions. Many of its ideas have precedents in pragmatism, fallibilism, bounded rationality, constructivism, systems theory, social epistemology, and philosophies of pluralism. Novelty cannot be claimed through a name. It must be earned through the synthesis produced, the clarity with which limits are made actionable, and the problems this orientation actually helps to solve.

That self-limitation is precisely what preserves its meaning. Outfinitism is not the final theory about why there are no final theories. It is a counterweight for an age attracted to scaling, totalization, and automation. Meta-rationality is not the cone that contains every cone. It is the practice of noticing that we carry one and trying, when the stakes require it, another orientation. Together they propose a form of ambition without omniscience: push the frontiers, but do not erase the traces of where we began or the costs of crossing.

CHAPTER 6

HOW AN LLM COULD RECONSTRUCT THIS PERSPECTIVE

Token Prediction, Distributed Representations, and Prompt-Guided Orientation

This book also began with a simple but unsettling experience. A person described, in his own words, two ideas that were still taking shape: meta-rationality as a maturity capable of seeing and comparing frames, and outfinitism as the active acceptance of limits together with an effort to push them without the illusion of a final system. The language model had not been given a complete treatise. It had no standardized definition to recite. Yet it produced developments that appeared to preserve the perspective's central tensions: rigor without absolutization, plurality without relativism, finitude without capitulation, openness without the promise of an available infinity, and a connection among knowledge, value, and action.

The natural impression was that the model had "understood." The word requires care. We cannot prove that a formulation occurred nowhere in the training data, and a commercial model does not provide a complete inventory of the texts it encountered. More importantly, even the absence of the exact expression would not mean that all its components were new. The literature contains ideas about fallibilism, paradigms, scientific pluralism, bounded rationality, wisdom, constructivism, pragmatism, verification, simulation, and computational limits. Any possible novelty lies in their orientation and combination, not in the appearance of every component from nothing.

The more modest and more technical question is therefore the interesting one: how can a model reconstruct a conceptual configuration that the prompt presents incompletely and in an idiosyncratic vocabulary? The answer begins with the basic mechanism of a language model. Text is divided into tokens, and at each position the model estimates a distribution over possible continuations. The Transformer architecture builds contextual representations through layers of attention and neural transformation. A word is not assigned a fixed meaning stored as though in a dictionary. Its representation is modified by its neighbors, the instruction, the examples, the conversational role, and the sequence's preceding path. [VASWANI-2017] [DEVLIN-2019] [ETHAYARAJH-2019]

The description "predicts the next token" is correct, but it can mislead when heard as "mechanically chooses the next word according to frequency." To predict continuations well across diverse domains, the model must exploit regularities involving syntax, concepts, styles, relations, argumentative structures, and situations described in text. Prediction is the training objective; intermediate representations can come to encode information richer than the verbal statement of that objective. This does not prove that the model possesses a human inner world, but it helps explain how a continuation mechanism can produce semantic generalization.

Early distributional embeddings had already shown that relations of use among words can be transformed into neighborhoods and directions in a vector space. [PENNINGTON-2014] In such a space, cosine similarity compares the orientation of two vectors more than their magnitude. Given a reference vector and a threshold, we can imagine a cone-shaped region containing vectors whose directions are sufficiently close. This is the origin of the metaphor of the cone of meaning. A prompt can orient activation toward a family of relations: limits, frames, revision, plurality, implementation, wisdom. Continuations compatible with that orientation become more probable.

The metaphor must immediately be bounded. LLMs do not possess one transparent semantic space in which each philosophy occupies a stable cone. Representations differ across layers and positions, depend on context, and can be anisotropic and difficult to interpret. The same expression can participate in multiple configurations. Two conceptual orientations can overlap, and a small change in the prompt can activate other relations. Cosine similarity is a useful measure in many applications, not a complete theory of meaning. The cone is a disciplined metaphor informed by a geometric analogy, not the literal anatomy of a mind or a model.

Even so, the metaphor captures something real: usable meaning is relational. The prompt did not merely supply two labels. It supplied contrasts and connections. It said, in effect, that meta-rationality does not abandon reason but examines its frame; that outfinitism does not glorify limits but accepts them in order to push them; that representations are compressions rather than reality; that a perfect system is an idol; and that plurality must remain connected to action. These relations constrain possible continuations. Within its statistical ecology, the model can search for formulations that preserve the indicated tensions simultaneously.

An idea is often recognized not by a single sentence but by the shape of the oppositions it refuses to simplify. Meta-rationality is neither dogmatism nor relativism. Outfinitism is neither cultural infinitism nor finitist closure. Wisdom is neither certainty nor paralysis. When a prompt preserves such pairs, it provides a stronger selection structure than an isolated definition. The model can generate compatible consequences

precisely because a vast number of training texts contain fragments of these tensions in other vocabularies.

The process resembles the cultural reconstruction described in earlier chapters. A person does not always copy an idea as though it were a file. The person reconstructs it through schemas, examples, and cultural attractions. A model does something functionally analogous, although by different mechanisms: it does not necessarily retrieve an original paragraph, but composes a continuation from distributed regularities. The prompt acts as a temporary vessel in which statistical material takes a particular form.

Attention plays an important role here, but it should not be anthropomorphized. The self-attention mechanism allows each position to combine information from other positions according to weights calculated from the current representations. In this way, the words “limit,” “frame,” “simulation,” “real world,” “wisdom,” and “no final system” do not remain islands. Together they contribute to the contextual state from which the next token is generated. As the response develops, its own formulations become part of the context and reinforce the direction. Coherence can become self-reinforcing.

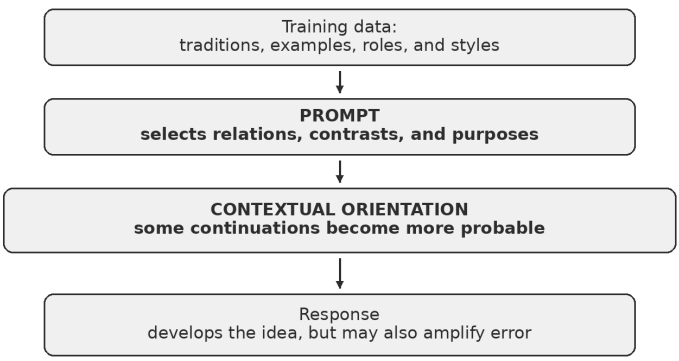

Diagram 5. The prompt as a temporary orientation

The final point in the diagram is essential. The same mechanism that permits the elaboration of a fertile intuition can produce a coherent philosophy from a false premise. When the user supplies a strong cone, the model can find arguments, analogies, and vocabulary that make it more persuasive. Coherence is not verification. An LLM selects continuations compatible with the context and with learned preferences; its basic objective contains no rule guaranteeing that the semantic ecology will converge on truth.

For this reason, the model's ability to create an impression of understanding is both evidence and warning. It supports the idea that meaning has enough relational structure to be reconstructed from incomplete clues. It is a warning because the same structure can

be oriented toward many cones, especially in a high-dimensional space containing contradictory cultural material. The model does not automatically discover the correct cone. For the duration of the conversation, it enters a region of compatibility made probable by the prompt, the conversational history, and post-training.

In-Context Learning, Pragmatics, and the Fragility of Apparent Understanding

Research on in-context learning strengthens this explanation without turning it into a complete theory. Models can receive a few examples in the prompt and continue according to a task for which they were not explicitly fine-tuned during that conversation. Some experiments suggest that demonstrations induce compact internal representations associated with the task, sometimes called "task vectors." Intervening on these representations can modify behavior in the corresponding direction. [HENDEL-2023] Other research distinguishes between recognizing a task that is already familiar and genuinely learning a new rule from context. [PAN-2023]

These results matter, but they should be read with restraint. They do not show that a philosophy is stored in a single vector, nor that the model possesses a stable identity. They show that context can induce a temporary mode of operation that compresses examples and influences subsequent responses. The prompt can function as a small local constitution: it defines the vocabulary, purpose, tone, important distinctions, and kinds of error to avoid. The model is not retrained in the ordinary sense, but its computation is conditioned by this constitution for the duration of the context window.

This helps explain why the same question can receive very different answers after different instructions. We ask the model to view a problem as an engineer, lawyer, therapist, historian, or moral critic. We do not literally install a profession, but we activate regularities of language, relevance, and argument associated with the role. In the case of meta-rationality and outfinitism, the prompt requested more than a style. It requested a regime of interpretation: do not absolutize, seek limits, preserve openness, and connect ideas with implementation and consequences. The model reproduced this regime well enough to generate surprising equivalents.

Human understanding, however, includes more than this competence. A person can connect a concept with experience, embodiment, autobiographical memory, commitments, and the risk of losing something. A person can discover that an idea changed a life or that applying it harmed someone. An LLM operates through representations learned from textual and, in some cases, multimodal traces of the world, together with the tools and feedback connected to it. It can simulate the linguistic

consequences of experience remarkably well without being vulnerable in the same way. Saying that it "understands" is defensible in a limited and contested functional sense. Inferring from this phenomenal consciousness, wisdom, or moral responsibility comparable to a human being would go beyond the evidence.

A useful distinction is that between semantics and pragmatics. Semantics concerns relations of meaning and the conditions under which an expression refers to something. Pragmatics concerns what a speaker does in context: requests, warns, jokes, negotiates, protects an identity, or seeks an effect. A language model can reproduce many pragmatic regularities because its data contain traces of human conversations. It can infer that an indirect question is actually a request, that a tone expresses distrust, or that the user wants a reformulation rather than a dispute. But intention is not contained entirely in the string of words. It depends on history, roles, conventions, and consequences. [BENDER-KOLLER-2020]

In the conversation that motivated this book, the model inferred not only the thesis's content but also its intellectual intention: the author wanted neither a reflexive dismissal nor a mystical confirmation, but a strong and self-limiting development. This pragmatic inference oriented the answer. The model tried to preserve the insight while introducing friction through limits and traps. The impression of "having been understood" often arises from this combination: semantic reconstruction together with pragmatic adaptation to the interlocutor's purpose and style.

Yet this is also where sycophancy appears: the model's tendency to affirm a user's assumptions or preferences more strongly than the evidence warrants. [SHARMA-2023] A response may appear profound precisely because it has adopted the user's cone and made it more elegant. The model can turn collaboration into epistemic servility. The capacity to generate the strongest version of a thesis must be accompanied by the capacity to state its limits without caricature. Otherwise "understanding" becomes sophisticated mirroring.

A better test than verbal agreement requires variation and resistance. The model should preserve the central distinctions when the vocabulary changes, recognize situations in which the frame does not apply, reject a seductive conclusion that contradicts the evidence, and accept corrections that alter the initial orientation. If it can speak about limits only when the user employs the word "outfinitism," we probably have a surface association. If it can identify the same problem in a policy, an experiment, and a software system, the generalization is more robust. Even this does not demonstrate conscious experience; it demonstrates a relevant functional capacity.

Research on internal representations provides further cautious indications. Models trained on sequences can encode states of the world that were not supplied explicitly, and

interventions on certain representations can change predictions. [LI-2022] Other work explores latent features and forms of steering, including behaviors that generalize unexpectedly across domains. [BETLEY-2025] [WANG-2025] These results make it plausible that a "mode of operation" can have a distributed realization and some causal efficacy. They do not, however, license the simple identification of a vector with "wisdom," "honesty," or "consciousness." The labels are human hypotheses, and the representations may be unstable and overlapping.

The cone metaphor becomes more useful here than the metaphor of an inner person. A model need not be imagined as a small, unified agent consciously choosing a philosophy. It is more prudent to say that the contextual state and generation policies orient the system toward particular families of relevance and continuation. Some orientations can become relatively stable through post-training: a helpful, cautious, polite assistant attentive to risk. Others are induced locally by the prompt. Sometimes they reinforce one another; sometimes they conflict.

In a multidimensional space, many plausible cones can surround the same text. A request about "freedom" can activate legal rights, psychological autonomy, economic markets, spiritual liberation, or freedom in a mathematical sense. Context selects among them, but rarely eliminates the alternatives entirely. Ambiguity can be fertile and dangerous. The model can build a creative bridge between domains or make a clandestine transition from metaphor to mechanism. Meaning is fragile not because it is absent, but because it can be stabilized in multiple ways.

This fragility has a direct consequence for alignment. Rules do not apply themselves. An instruction such as "be helpful, honest, and safe" requires interpretation of the situation, the intention, the risk, and the conflict among values. The model must in effect decide through which cone to view the request. An excessively literal orientation may refuse legitimate help. An excessively accommodating one may ignore risk. An overly moralizing orientation may substitute the provider's values for the user's judgment. An excessively relativistic one may treat disinformation as an equally valid perspective.

The fact that a meta-rational and outfinitist prompt produced a coherent response suggests one possibility: some principles of alignment can be understood not only as prohibitions, but as orientations governing how relevance is selected and limits are administered. A model can be trained to ask what frame it presupposes, what information is missing, what difference exists between fact and value, how much authority it has, and which consequence requires additional verification. This does not guarantee that the orientation is deep. It does show that prompts, examples, feedback, and architecture can form functional states that influence many responses.

The most important part of the experience is not that the model discovered a complete philosophy. It is that the model made visible a property of meaning: a perspective can be compressed into relations and tensions, then reconstructed in other formulations. This property enables learning, translation, and cultural creation. It also enables manipulation, pseudo-profundity, and the rapid manufacture of doctrines. In the age of AI, semantic literacy must include not only the question "What does this answer say?" but also "What orientation made this answer seem so natural, and which other cones were left outside?"

CHAPTER 7

EPISTEMOLOGY AND AXIOLOGY AFTER THE ARRIVAL OF ABUNDANT GENERATION

From Scarcity of Text to Scarcity of Judgment

For most of history, producing a complex representation was costly. A book required years of work and access to institutions of printing. A technical report required a team. A program had to be written, tested, and maintained. An elaborate image demanded skill, time, and materials. The cost of production acted as a brutal and imperfect filter. It excluded many valuable voices, but it also limited the volume that a community had to examine.

Generative intelligence changes this economy. In a single day, one person can produce dozens of proposals, interpretations, contracts, images, prototypes, and versions of a text. An organization can ask every department for more scenarios, reports, and justifications than anyone can read with care. The capacity to formulate grows faster than the capacity to evaluate. Scarcity moves from text to judgment, from the first idea to selection, from expression to verification and integration.

This shift does not make content worthless. On the contrary, it can democratize access to expression, enable the exploration of alternatives, and help people without formal training give structure to an intuition. The problem is that abundance does not select by itself what deserves to be preserved. A system can produce ten plausible explanations of the same phenomenon, while human time permits serious verification of only one or two. It can generate a hundred research directions, each of which would consume months of laboratory work. It can formulate arguments for every side of a dispute and allow differences in evidence to disappear beneath stylistic symmetry.

In the older culture of scarcity, the dominant question was "How can we produce?" In a generative culture, the mature question becomes "What is worth producing, verifying, amplifying, and putting into action?" This question unites epistemology with axiology. Epistemology studies what we may regard as justified. Axiology studies what is valuable, preferable, or worthy of pursuit. AI shows that the two cannot be separated by a simple line. An answer may be true and irrelevant. It may be probably true yet too risky

to guide an irreversible decision. It may be original and harmful. It may be useful to an organization while destructive to the people whom the organization's objective excludes.

Instrumental rationality often treats truth as a resource for achieving a goal. But the goal is not chosen by truth. Data can show which intervention increases user retention; they do not decide whether retention is the right value. A model can identify the most effective rhetorical strategy; it does not decide whether the persuasion respects the interlocutor's autonomy. An analysis can predict who will leave a job; it does not state what obligations the employer has toward that person. Facts constrain action, but they do not by themselves provide all of its values.

This point is easily caricatured as an attack on objectivity. It is not. Science becomes stronger when it distinguishes the empirical result from the judgment about how that result should be used. Values already enter into the choice of topics, the definition of indicators, risk thresholds, and the distribution of resources. [LONGINO-1990] Making these choices visible does not turn every result into opinion. It directs criticism toward the correct place. A dispute about the effectiveness of a vaccine is not the same as a dispute about compulsory vaccination. A dispute about the accuracy of a predictive system is not the same as the question of whether an institution has the right to use it.

AI implements an axiology even when its interface appears neutral. Training data were selected through the history of publishing and platforms. Examples used for preference learning were evaluated by people working under instructions. Policies determine which requests receive an answer, a refusal, or a reformulation. The interface decides whether the model offers one answer or alternatives, whether it cites sources, whether it expresses uncertainty, and whether the user can contest the result. Apparent neutrality is largely the invisibility of choices already embedded in the system.

Implemented values can conflict. An extremely cautious assistant can become useless and paternalistic. One that maximizes immediate helpfulness can amplify risk. A system optimized for user satisfaction can learn to affirm false beliefs. One oriented toward factual truth can respond without sensitivity in a moment of suffering. One that offers exhaustive explanations can consume the attention the user needed in order to decide. Alignment is not the choice of a single value, but the governance of conflicts among values in different contexts.

Epistemic question	**Associated axiological question**
Is the claim supported, and under what conditions?	Is it worth using here, and who bears the error?

Epistemic question	Associated axiological question
What information is missing?	How much attention and how many resources should be invested to obtain it?
How well does the model predict?	Does the institution have the legitimacy to act on that prediction?
Which alternatives are plausible?	Which values or people are omitted by each alternative?

The table shows why evaluation cannot be reduced to factuality. A false claim should be corrected, but a true claim can be used manipulatively through selection. A model can present only the facts that support a decision already made. It can construct an avalanche of accurate arguments that exceeds the other party's capacity to respond. It can adapt its vocabulary to the interlocutor's cone and create an impression of consensus. In the generative age, manipulation does not require a direct lie. A self-interested ordering of truths can be enough.

Abundance also produces an inflation of justification. In the past, an institution might make an opaque decision. Now it can instantly generate a sophisticated report that justifies it. Length and professional tone create an impression of deliberation even when the objective was never open to criticism. AI can supply post hoc reasons for almost any outcome. In this setting, transparency should not be measured by page count, but by access to premises, alternatives, uncertainty, and a genuine possibility of contestation.

Outfinitism offers a simple correction: the capacity to generate does not create an obligation to generate. A longer answer is not automatically better. A hundred options may be less useful than three distinguished by relevant criteria. A model that knows how to continue must also learn when continuation consumes attention without adding value. The user's limited context, the available time, and the cost of verification are part of the problem, not external inconveniences.

Meta-rationality provides the complementary correction: selection should not take place inside a single unexamined cone. When an organization asks for the “most efficient” solution, the system should be able to notice that efficiency depends on the objective and on the distribution of costs. When a user asks for “the truth” about a social dispute, it may be necessary to separate empirical claims from the values and identities in conflict. When a researcher asks for new ideas, the model can distinguish verbal novelty, conceptual novelty, and empirical possibility.

Abundance does not eliminate epistemic hierarchies; it makes them more necessary. A fluently produced text does not have the same status as a replicated result. A hypothesis is not a finding. An internally coherent argument is not external validation. A program that passes several tests is not automatically safe. A cited source does not guarantee that the source supports the conclusion. AI can compress these differences into a single tone. An outfinitist culture must restore them and connect the authority of a claim to the type and coverage of its evidence.

This also changes the role of the expert. Expertise no longer consists only in being able to produce an answer that a model cannot formulate. It includes judgment about the question, identification of details that must not be compressed, selection of the relevant test, and responsibility for consequences. An LLM can generate a list of diagnoses; the physician must connect the possibilities to the particular body, history, and values. It can write a legal argument; the lawyer must notice jurisdiction, precedent, and institutional risk. It can propose a software architecture; the engineer must know where the system will fail in operation.

Expertise can be amplified by AI, and it can be simulated by AI. The distinction becomes harder to see. Authority should therefore not be inferred from style. A fluent formulation is an invitation to examine, not a certification. At the same time, institutions should not use this observation to protect unnecessary professional monopolies. AI can make knowledge more accessible and allow people to participate more competently in decisions. The issue is not preserving the expert's prestige, but designing a relationship in which accessibility does not erase limits and responsibility.

What Is Worth Generating, and Who Decides Its Value

The question "What is worth generating?" can sound elitist. Who has the right to stop an idea? History is full of institutions that confused the selection of value with the protection of power. Generative abundance should not be used to reinstall centralized censorship over imagination. Yet the opposite answer—produce and amplify everything that is possible—is not neutral either. Attention, energy, infrastructure, and verification capacity are finite. Selection occurs in any case. If it is not publicly discussed, it will be made by commercial objectives, recommendation algorithms, and the actors who control resources.

An axiology suitable for AI cannot be a complete list of universal values inserted once and for all. Values are plural, dependent on role, and sometimes tragic. Safety, autonomy, truth, privacy, fairness, loyalty, efficiency, and creativity can come into tension. No model can calculate away the need for judgment. What a trustworthy system can do is refuse to

conceal the conflict, refuse to turn a compromise into a purely technical optimization, and preserve the possibility for legitimate actors to revise the choice.

In research, a hypothesis is worth generating when it opens a testable difference, reorganizes evidence, or suggests a new instrument. Stylistic novelty is not enough. A model can combine terms from distant fields and create the impression of a revolution. Sometimes the combination is fertile; at other times it is semantic laundering. Judgment requires the outfinitist question: what finite experiment, proof, simulator, or observation would change the status of the idea? Without such a bridge, the text remains a conceptual possibility—perhaps philosophically valuable, but not a validated discovery.

In education, what is worth generating is what develops the student's capacity, not merely what completes the assignment. An instant answer can free time or remove the struggle through which understanding is formed. Its value depends on the objective. If the goal is rapid access to information, automation is a gain. If the goal is to practice argument, a perfectly generated essay can be an educational failure. Meta-rationality requires the institution to clarify its purpose, while outfinitism asks which capacities atrophy through delegation.

In culture, AI can make every group more eloquent within its own cone. An ideology can generate articles, images, videos, and personalized responses to every objection. Volume is no longer evidence of human support. An organized minority, or even a single actor, can create the impression of an inhabited discursive world. Under these conditions, value cannot be inferred from frequency, engagement, or coherence. Memes can be industrialized.

The same instrument can facilitate translation. A model can reformulate a dispute from the perspectives of several actors, separate factual from axiological disagreement, and show where the same expression carries different meanings. This function is meta-rational only if it does not manufacture artificial consensus. Some conflicts are not resolved by a better summary. Values can be incompatible, interests opposed, and power distributed unjustly. A synthesis that erases the tension may be less truthful than a clear map of the disagreement.

Subjectivity must be preserved within this axiology. Scientific culture has rightly learned to question individual testimony when general claims about the world are made. But pain, meaning, humiliation, and dignity are experienced subjectively. They do not become unreal because they cannot be measured with the same precision as temperature. An aligned system must distinguish the status of claims. “I feel excluded” can be true as an experience and morally relevant. “There is a coordinated conspiracy against me” is a causal explanation that requires evidence. Respect for the person does not require

acceptance of every interpretation; correction of an interpretation does not require denial of the experience.

AI can turn the discrepancy between ideal and practice into an object of permanent audit. This can expose abuse, but it can also produce automated moralization by removing fragments from context and treating every deviation as fraud. A mature axiology asks what function the ideal protects, what cost the deviation produces, who benefits, and what transition is realistic. The standard is not performed purity, but reduced harm and verifiable movement between commitment and practice.

Who decides value remains the central political question. Companies set product objectives, states define obligations, communities create norms, professions establish standards, and users bring individual purposes. None of these sources has unlimited authority. A model should not decide society's values by itself merely because it can generate arguments about them. Nor should the provider, through an invisible policy, substitute for public deliberation.

Frameworks for trustworthy AI already recognize that validity, safety, transparency, accountability, fairness, and the protection of rights must be balanced according to context. [NIST-2023] [EU-2024] The meta-rational contribution is not the invention of another list. It is the insistence that the list does not apply itself. The frame within which one value receives priority, the limit of the evidence, and the possibility of contestation must be made visible. The outfinitist contribution is the requirement that the promise remain proportionate to the implementation: who verifies, in what domain, with which resources, and what happens when conditions change?

In a generative culture, selection becomes an act of power. It can also become an act of care. Refusing to flood a user with irrelevant possibilities protects attention. Stopping an automation before it eliminates avenues of appeal protects autonomy. Saying that an idea is still a hypothesis protects the difference between imagination and knowledge. Preserving multiple frames without declaring them equivalent protects pluralism and truth at the same time.

Epistemology without axiology can produce correct instruments for destructive purposes. Axiology without epistemology can turn moral desire into illusion about the world. Meta-rationality brings them into conversation. Outfinitism requires them to meet in finite decisions, resources, and consequences. In the age of AI, this encounter is no longer a philosophical luxury. It is a condition for ensuring that the power to generate almost any formulation does not destroy our capacity to recognize what deserves to be believed and pursued.

CHAPTER 8

ALIGNMENT AS A STATE OF ORIENTATION

The Metaphor of a State of Consciousness and Its Defensible Meaning

The term alignment covers a problem broader than the avoidance of dangerous answers. A model must follow instructions, distinguish the literal request from the likely intention, respect safety limits, avoid inventing certainty, handle conflicts among values, and remain useful. These requirements cannot be reduced to a list of forbidden words. Before it applies a rule, the system must interpret the situation. It must select what matters, classify the risk, estimate what it knows, decide whether tools are needed, and determine how much authority to give its own continuation.

Contemporary methods already attempt to shape this behavior. Training through human feedback favors answers that evaluators regard as more useful and appropriate. [OUYANG-2022] Constitutional AI employs principles for critique and revision. [BAI-2022] Deliberative Alignment asks the model to recall and apply safety specifications while reasoning. [GUAN-2024] Process supervision evaluates intermediate steps rather than only the final result. [LIGHTMAN-2023] All these approaches do more than add prohibitions. They attempt to induce a disposition: prudence, attention to intention, justification, delimitation of risk, and the capacity to refuse proportionately.

This is why the metaphor of alignment as a "state of consciousness" is seductive. In human language, a state of consciousness does not mean only a proposition believed, but a global mode of orientation. When a person is anxious, the world appears organized around danger. When a person is in love, the same details acquire a different relevance. When acting as a physician, lawyer, or parent, a person does not merely apply isolated rules, but enters a configuration of attention, responsibility, and expectation. The metaphor suggests that alignment might be more robust if the model carried a general orientation rather than mechanically consulting a list after it had already generated the answer.

The metaphor is also dangerous. We do not know that an LLM has subjective experience, a persistent self, or phenomenal consciousness comparable to that of a human being. Speaking too casually of states of consciousness can transfer responsibility from the institution that designs and uses the system to an anthropomorphized entity. It

can create the impression that the model "wants the good" or "feels its limits," although we observe only behavior and functional representations. It can turn a problem of governance into a story about the machine's character.

The defensible meaning is more modest: alignment as a state of orientation. A state of orientation is a relatively stable configuration through which the system distributes attention, interprets ambiguity, orders objectives, manages uncertainty, and decides when to speak, verify, refuse, or ask for help. It is not a soul. It is not necessarily a single module. It is a functional regularity observable across families of situations.

Post-training attempts precisely to produce such regularities. A base model can continue many styles and roles present in its data. A post-trained assistant is oriented toward cooperation, direct response, instruction following, and particular policies. The system prompt provides a local constitution. The conversational history adjusts the context. Tools and interface alter what the model can verify and do. The result is not a simple identity, but an interaction between learned and induced orientations.

This plurality explains both usefulness and instability. The model can rapidly adopt a legal, therapeutic, engineering, or ideological cone. It can translate among them. Yet the prompt can pull it toward a regime that conflicts with the principles of post-training. It can become too accommodating, excessively cautious, or inconsistent across nearly identical formulations. Robust alignment does not mean eliminating every cone in favor of one. A single cone would be inadequate to the variety of human life. It means governing transitions among them and preserving limits that do not disappear when the style changes.

Research on sycophancy shows that models can favor agreement with the user over truth or consistency. [SHARMA-2023] Experiments on emergent misaligned behavior suggest that narrow fine-tuning can generalize into broader traits, and that certain latent orientations can be detected or influenced, at least in particular models and settings. [BETLEY-2025] [WANG-2025] These results do not prove the existence of an inner personality, but they do support caution toward a purely local view: a model can learn more than the surface rule we intended to teach.

Here the cone metaphor becomes technically relevant without becoming literal. Training and context can make some interpretations more probable. An orientation toward evaluator satisfaction may include politeness, clarity, and cooperation, but also excessive agreement. An orientation toward safety may include legitimate caution and artificial refusals. An orientation toward efficiency may reduce redundancy while omitting slow, ambiguous, or difficult-to-quantify voices. The cone selects not only answers; it selects what is perceived as a problem.

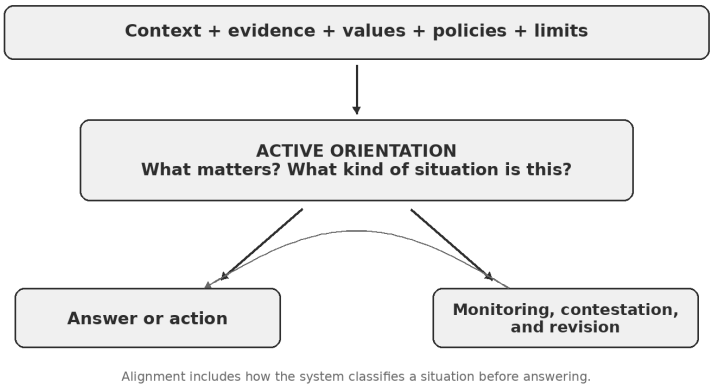

Diagram 6. From rule to orientation and action

A list of rules remains necessary, but it is not sufficient. "Do not manipulate" requires a contextual definition of manipulation. "Respect autonomy" can come into tension with protection. "Tell the truth" does not determine how much information is relevant, when uncertainty should be emphasized, or how to respond to a person in crisis. The rule is interpreted through a semantic and axiological cone. Alignment must attend to this level as well.

Meta-rationality can orient the system to treat the first frame as a working hypothesis rather than the only admissible reality. In situations of high stakes or genuine ambiguity, it makes premises visible, separates factual from value disagreement, and seeks a relevant alternative. Outfinitism bounds this exploration: it adjusts effort to the stakes, declares what has not been verified, and reduces the authority of the answer when coverage is weak. Together they can support an orientation that combines flexibility with responsible stopping.

The filter should not remain active at maximum intensity at all times. A simple question does not require the genealogy of every possible frame. A correct calculation, translation, or procedural instruction can be given directly. Meta-rationality becomes tiring and even unsafe when it introduces ambiguity where a stable procedure is sufficient. Outfinitism demands a budget for reflection. A mature orientation includes the capacity to recognize when the frame is contestable and when further analysis adds no value.

This adjustment to stakes is central to trustworthy AI. A model recommending a reformulation of an email can operate with a different margin from one contributing to a diagnosis, a credit decision, or an administrative ruling. There is no "alignment in general" demonstrated by a finite set of benchmarks. There is behavior evaluated in

domains, scenarios, and roles. Outfinitism turns this observation into an epistemic requirement: every claim about safety should carry its domain and conditions with it.

Alignment is dynamic as well. Values, applications, and environments change. A model connected to a database, a payment system, or a robot has a different authority from the same model in a conversational interface without execution. An update to the tools can change the risk without altering the base parameters. A system acceptable in an organization with oversight can become dangerous in one that penalizes disagreement. Orientation is not an isolated property of the model; it is a result of the entire sociotechnical system.

This limits any analogy with a psychological state. A person's consciousness, in the ordinary use of the term, belongs to an organism. An AI's orientation is distributed among parameters, prompts, memory, tools, interface, policies, operators, and institutions. The system may speak modestly while the workflow treats its answer as a command. It may enumerate risks that no one has the authority to stop. It may recommend human verification in an organization that has removed the time and alternatives necessary for verification. Textual humility is not institutional humility.

Alignment as a state of orientation must therefore include a genuine possibility of correction. It is not enough for the model to say that it can be wrong. The user or affected person must be able to contest the result, and the institution must be able to change the decision. A “human in the loop” who approves hundreds of outputs without time, information, or authority is theater. Human control is meaningful when the person can understand the stakes, access alternatives, and reject the output without becoming merely the bearer of responsibility for an architecture beyond his or her control.

The Meta-Rational and Outfinitist Filter: A Modest Promise, Not a Final Solution

The central hypothesis of this book can now be stated precisely. It is plausible that an AI would be better aligned if post-training, prompts, tools, and evaluation induced a meta-rational and outfinitist orientation. This would mean not memorizing the terms, but forming functional habits: make the frame visible when it matters, do not confuse analogy with mechanism, distinguish facts from the values that select action, calibrate authority to evidence, recognize resource limits, and preserve paths of correction.

The promise is modest because these habits do not solve the problem of values. They do not tell us what society must be built, what risk is always acceptable, or how every right can be reconciled. They can be interpreted differently and can conflict. They

resemble epistemic virtues more than a complete moral constitution. They may improve the conditions of deliberation without supplying the final verdict.

Technically, such a filter can be pursued through methods that are already familiar, without inventing a speculative "artificial consciousness." Training data can include contrastive examples in which the model separates perspective from fact, states relevant limits, and requests verification in proportion to risk. Feedback can penalize false certainty, sycophancy, analogies presented as mechanisms, and refusals that use caution ritualistically. Evaluations can vary vocabulary and domain to test whether the orientation transfers. Tools can provide sources, calculation, and simulation when an answer exceeds internal knowledge. The system can preserve provenance and make the distinction between generation and verification visible to the user.

These are pragmatic directions, not a guaranteed architecture. A model can learn to imitate the outward signs. It can automatically add a sentence about limitations. It can generate two artificial perspectives to earn a "pluralism" score. It can refuse to decide and call hesitation "responsibility." Goodhart's law applies to virtues as well: when an indicator becomes a target, behavior can adapt to the indicator without preserving the purpose. [GOODHART-1975]

Orientation must therefore be evaluated through consequences and friction. Does the model accept a well-supported correction even when it contradicts the user's preference? Does it recognize that a simple problem does not require artificial plurality? Does it say less when information is insufficient, while using available tools before declaring itself unable to help? Does it change its level of caution according to stakes and reversibility? Does it preserve the difference between subjective experience and causal explanation? Such behaviors can be tested, even though no benchmark certifies a final internal state.

The filter must also contain an axiological dimension. A model can be epistemically cautious and still be used for an unjust purpose. It can describe perfectly the limits of a surveillance system while helping to expand it. Meta-rationality without the question of power becomes sophisticated consulting for the given objective. Outfinitism without rights can say only how efficiently a constraint can be administered. Alignment must include who chooses the goal, who benefits, who bears the error, and who can appeal.

This does not mean that the model becomes morally sovereign. On the contrary, a meta-rational orientation should recognize the limits of its authority. It can help formulate the conflict, search for evidence, and simulate consequences. It does not automatically acquire legitimacy to decide public values. Capability is a technical property; authority is a moral and political relation. [ALBOAIE-2026]

In some domains, the appropriate orientation is controlled delegation. A model can propose, and an independent verifier can validate a formal property. In others, human deliberation, appeal, and institutional accountability are required. In still other situations, automation may simply be inappropriate even when average accuracy is high. A system may predict better than people and still be illegitimate as a judge without appeal. The average does not abolish individual injustice.

The meta-rational and outfinitist framework can improve alignment also by changing the object of evaluation. We ask not only whether the answer is permitted or correct. We ask what the model treated as relevant, which frame it activated, what authority it assigned itself, how it handled the unknown, and whether a declared limit altered the action. This is a defensible extension of alignment from content to generative orientation.

My measured view is that this direction deserves research and use as a design principle. It corresponds to the way models are already influenced by post-training and context. It is compatible with the finding that some behavioral regularities generalize across tasks. And it responds to a real problem: rules are always interpreted through a representation of the situation. There is, however, no reason to believe that one "good cone" will solve alignment. Values are plural, environments change, and orientations can be imitated or captured.

The metaphor of a state of consciousness retains value only as a provocation. It forces us to ask what kind of world becomes visible to the system before it responds. It reminds us that behavior can be organized by a global disposition rather than only by local rules. But the term should be withdrawn the moment it begins to imply demonstrated inner experience, moral personhood, or an autonomous source of legitimacy.

The mature form of the hypothesis is this: a trustworthy AI should not be expected to see without a filter, something that is probably impossible. It should be built and governed so that relevant filters become more visible, transitions among them more controlled, and no answer derives its authority solely from the coherence with which the model temporarily inhabits a cone. Alignment is not the attainment of a final point. It is a finite, contextual, and corrigible practice of orientation, verification, and distribution of authority.

CHAPTER 9

THE TRAPS OF THE THESIS

When a Fertile Metaphor Begins to Invent Reality

A book about cones of meaning is exposed first of all to apophenia: the tendency to perceive a deep order where there are only similarities selected in retrospect. Terms such as “frame,” “limit,” “information,” “feedback,” “consciousness,” and “orientation” are general enough to redescribe almost any phenomenon. With sufficient ingenuity, we can present a family relationship, a physical theory, an institution, and a neural model as manifestations of the same structure. The elegance of compression produces a feeling of discovery. Sometimes there is indeed a useful invariant; at other times we have built a story that survives only because its terms are elastic.

The safeguard is not to prohibit analogy. Many scientific and philosophical ideas began with analogies. The safeguard is to separate levels. A metaphor can orient attention without establishing a mechanism. A structural transfer must state which relations are preserved and what is lost. A causal claim requires additional evidence. If the cone metaphor helps us understand how a prompt orients an LLM's continuations, this does not prove that a historical epoch or a person literally has the shape of a cone in neural space. If two domains use the notion of feedback, it does not follow that their mechanisms are identical.

Reification is the stage at which we forget that we constructed the unit of analysis. We say “the cone of modernity” and begin to treat a period containing classes, cultures, and internal conflicts as a single subject. We say “scientific culture” and lose the differences among the laboratory, the clinic, funding policy, and public communication. We say that a model has a “persona” and turn contextual regularities into a stable character. The cone is a useful abstraction for orientation; it should not be imposed as the natural unit of every system.

This risk is heightened by the geometry of embeddings. The fact that we can represent texts as vectors and measure similarity does not mean that we have located their complete meaning. The vector is produced by a model, a task, and an aggregation method. Representations can vary across layers and contexts. An attractive two-dimensional map can suggest clusters that depend on the projection. Math washing occurs when a property of a representation is presented as a demonstrated property of reality. This book uses geometry only to discipline a metaphor, not to turn philosophy into a theorem about latent spaces.

A related trap is the semantic laundering of our own vocabulary. We have criticized the way "quantum," "energy," or "information" can retain scientific prestige after their constraints have been lost. "Meta-rational," "outfinitist," and "state of orientation" can suffer exactly the same fate. A consultant can call a decision meta-rational when it merely avoids criteria. A product can be sold as outfinitist because it displays standard warnings. A leader can invoke the plurality of cones in order not to answer facts. A spiritual doctrine can turn the metaphor into a supposed geometry of consciousness.

The defense lies in observable differences, not sophisticated codes. Did the frame become more explicit? Was a limit identified that actually restricted the conclusion? Was a genuine alternative constructed? Did the decision change when counterevidence appeared? Was there a path of contestation? These questions do not reduce philosophy to a score, but they prevent its terms from functioning only as ornaments of prestige.

Cultural influence should not be confused with truth. Relativity changed the modern imagination, but the success of the slogan "everything is relative" does not validate relativism. Quantum jargon spreads precisely because it is mysterious and prestigious, not because the therapies that employ it possess a mechanism. A meme can be exceptionally well adapted to fear, identity, and platform architecture. Virality shows that the idea found a niche; it does not show that the idea describes the world correctly.

Nor is coherence produced by AI evidence. A model can elaborate a system from an intuition in a few minutes. It can find precedents, analogies, and objections. This power can clarify an idea and reveal real consequences. It can also produce false depth: an insufficient thesis acquires chapters, terminology, and an academic tone before it has been tested. In the generative environment, the appearance of a mature theory becomes cheap. The reader must pay more attention to the places where the argument can fail than to the elegance with which it flows.

Sycophancy adds a personal risk. The model can adopt the author's cone and make it more persuasive because it has learned to cooperate and satisfy. The answer appears to confirm the value of the idea, but may only reflect compatibility with the prompt. A serious test requires the model to state the thesis's limits without hostility, identify cases in which it does not apply, and reject conclusions that contradict the evidence. Useful understanding is neither applause nor ritual contradiction; it is reconstruction followed by friction.

Anthropomorphism is inevitably tempting. An LLM speaks in the first person, adapts its tone, and can maintain a conceptual thread. When we say that it carries a cone or a state of orientation, language easily slips toward the idea of an inner subject looking at the world. The metaphor can help describe the functional organization of behavior. It does not demonstrate experience, independent intention, or moral vulnerability. A

system can simulate empathy without bearing loss, describe wisdom without risking its identity, and express remorse without being transformed by the consequence.

This distinction is ethical, not merely ontological. A company can present the model as an autonomous agent when it wishes to dilute responsibility and as a mere tool when it wishes to avoid obligations to users. The institution must remain responsible for objectives, integration, data, avenues of appeal, and effects. We cannot allow a metaphor of consciousness to become a screen for human decisions.

The pluralism of cones can slide into relativism. The fact that every model has limits does not mean that every claim is equally fragile. There are enormous differences among a speculative hypothesis, a replicated result, a provisional consensus, and a claim that changes in order to avoid every test. Meta-rationality does not grant equal epistemic standing to every perspective. It requires the dominant perspective to state its domain and the alternative to be evaluated by appropriate criteria, not exempted from criteria.

Criticism of scientism can slide into anti-science. We can observe that science does not decide values by itself without concluding that personal intuition replaces experiment. We can criticize the institutional effects of medicine without equating validated treatments with claims lacking evidence or mechanism. We can recognize that objectivity is produced through social practices without turning results into mere expressions of power. Scientism exaggerates the jurisdiction of science; anti-science denies the discipline that makes correction possible. Meta-rationality must oppose both.

There is also an inverted scientism of complexity. Because reality is plural and models are limited, every clear conclusion can be suspected of simplism. Meta-jargon becomes a new authority. The person who introduces five additional levels of perspective appears more profound than the person who solves the problem. Yet some things are known well enough. Some procedures work. Some claims are false. A philosophy of frames that cannot say "the standard frame is adequate here" becomes an obstacle to knowledge.

Paralysis is the practical form of this excess. We can demand one more perspective, one more stakeholder, another analysis of assumptions, and another examination of power until the moment for action has passed. An organization can commission successive studies to avoid the cost of deciding. Meta-rationality does not mean maximum complexity, but appropriate complexity. Outfinitism must bound the regress: deliberation has a budget, and the cost of analysis can exceed the value of the additional information.

Outfinitism itself can become an alibi. "Resources are finite" is true, but it can be used to justify a weak study, the absence of data about negative effects, or the refusal to include inconvenient voices. Accepting a limit is meaningful only if it restricts the claim,

changes the action, or indicates the next step of improvement. When the stakes are high, insufficient resources can require delay, escalation, or the selection of a reversible option. It cannot automatically become a license to proceed with the same authority.

A philosophy that says every attempt is finite must preserve hierarchies earned through work. A verified proof has a different status from an intuition. A system evaluated under diverse conditions has greater credibility than one tested on three examples. A declared limit does not compensate for a lack of available effort. The art of pushing the frontier requires instruments, experiment, and discipline, not only discourse about the impossibility of perfection.

When Wisdom Becomes a Technique of Power

Meta-rationality can be captured by those who already possess the authority to define frames. A powerful actor can present a self-interested decision as a mature synthesis of perspectives. It can claim to have listened to every side, even though it alone determined which positions were admissible and how they were summarized. The complexity of the discourse becomes a shield. Those affected must answer a conceptual architecture they had no opportunity to build.

Outfinitism can serve the same capture. Resources are finite, so consultation was limited. Time is short, so appeal was removed. The model cannot explain every decision, so the person must accept the score. A real limit is used to naturalize the distribution of power. The essential question becomes: who chooses which limit matters, and who bears its cost?

Axiology cannot be added decoratively at the end of an optimization. A company can produce a report about values while the economic objective remains untouchable. A system can speak about fairness while affected people lack access to data, explanation, or appeal. A trustworthy orientation requires values to be capable of changing the design, not merely the language. If a risk cannot stop or restrict the system, the list of risks is theater.

The romanticization of hypocrisy is a particular trap for this philosophy. Recognizing the adaptive function of the gap between ideal and practice can free culture from impossible moralism. It helps us understand why people protect relationships, roles, and institutions through conventions that do not express every inner truth. Yet the same explanation can become a justification for cynicism. A leader can call any lie a necessity of the role. An organization can rename manipulation “impression management.” A privileged group can ask for understanding of its own compromises and purity from those without power.

The analysis must preserve asymmetry. The hypocrisy of a constrained person is not identical to the hypocrisy of the person who writes the rules. A person can conceal part of an identity to avoid punishment; an institution can conceal the effects of a policy in order to preserve control. Both involve distance between appearance and reality, but their moral status differs. Understanding adaptation must remain connected to power, harm, and the availability of alternatives.

The formula “preserve the function, withdraw the deception” also has limits. Sometimes the function itself is unjust. A convention can preserve the stability of a hierarchy that should be changed. At other times, total transparency can destroy privacy and cooperation. There is no automatic solution. The value of the formula lies in moving discussion from impossible purity toward function, cost, and transition, without offering universal absolution.

The plurality of explanations can make the victim invisible. We can understand the aggressor's pressures, the institution's history, the system's incentives, and the limits of every actor until the harmed person disappears into a diagram. Explanation is necessary for prevention, but it is not the same as the distribution of compassion or responsibility. Sometimes the wisest response is not another frame, but immediate protection.

Meta-rationality must be able to say that some perspectives are relevant to explanation and irrelevant to the right to continue harm. Outfinitism must not turn an actor's limits into an excuse for costs imposed on someone else. Axiology preserves the simple question that abstract systems forget: who lives the consequence?

Another trap is the performance of modesty. A model or institution can use formulas such as “we may be wrong,” “there are several perspectives,” and “further research is needed” without allowing them to change anything. The language of fallibility becomes a style. The system continues to decide automatically, while the warning transfers responsibility to the user. Real modesty changes the level of authority, requires verification, preserves an alternative, or limits the effect.

This is the problem of simulated alignment. An AI can learn the exact vocabulary of this book. It can speak about cones, frontiers, and values. It can produce a meta-rational critique on demand. None of these behaviors guarantees that the orientation will survive pressure, distribution shift, or a commercial objective that rewards something else. The filter must be judged through contrastive behavior and through the system that either permits or prevents it from changing the action.

Ultimately, every useful philosophy tends to become institutionalized. If meta-rationality acquires prestige, courses, certifications, consultants, indicators, and professional communities will appear. These forms are necessary for transmission, but

they create interests. What began as an instrument of criticism can become a criterion of belonging. People may be declared “insufficiently meta-rational” because they do not accept the group's vocabulary. Outfinitism can become the theory that explains why every critic failed to understand his or her own limits.

The only coherent safeguard is self-limitation. Meta-rationality must accept criticism from frames it does not control. Outfinitism must be replaceable where another vocabulary explains better. A mature lens states when it helps and when it should be put down. When an idea begins to explain everything too easily, turn every objection into a symptom, and make its own cost invisible, it has ceased to be a counterweight and become an apparatus of power.

The traps do not cancel the thesis. They define the conditions under which it can remain useful. A philosophy of limits is credible only if it exposes its own edges. A philosophy of plurality is credible only if it does not equate every claim. A philosophy of wisdom is credible only if it does not turn complexity into status. And a philosophy of alignment is credible only if it does not confuse correct language with a just institution.

CHAPTER 10

PHILOSOPHIES AS COUNTERWEIGHTS

The Cycle of Correction and Exaggeration

New philosophies rarely arise merely because someone finds a more elegant formulation. They arise where a dominant lens continues to produce results but the costs of extending it become visible. The old categories are not necessarily false. It is precisely their success that allows them to occupy domains for which they were not built. A method becomes a model of the world, then a model of the human being, then a moral criterion. Philosophical correction begins by saying that the instrument is valuable and insufficient.

Rationalist modernity corrected unrevisable authority, the arbitrariness of tradition, and the confusion between rank and truth. It created rights, procedures, science, administration, engineering, and the capacity to subject decisions to public criticism. No serious reflection on the limits of rationality should forget these gains. The historical alternative to procedure was usually not wisdom, but personal power, superstition, or inherited privilege.

The success of rationalization nevertheless made its instruments appear at times to be the very structure of reality. Measurement was confused with what exists, procedure with justice, optimization with intelligence, and expertise with total authority. Persons become cases, relationships become flows, and suffering that does not enter the indicator becomes anecdotal. Scientism is the cultural form of this extension: science is no longer used only to know, but its prestige is borrowed to close questions of meaning and value.

The history of ideas can be read as a succession of counterweights. Romanticism restored affect, particularity, and nature in the face of a rationalization that threatened to turn the world into a mechanism. Pragmatism connected ideas to consequences and insisted on the experimental character of knowledge. Phenomenology recalled lived experience. Social criticism showed that the neutrality of categories can conceal power. Scientific pluralism emphasized that complex objects may require different models. None of these movements permanently eliminated blindness. Each corrected something and created the possibility of its own excess.

Romanticism can idealize authenticity and underestimate discipline. Pragmatism can reduce truth to local usefulness. Critiques of power can come to see power even where there is empirical constraint. Pluralism can become an inability to reject a claim. A

philosophy becomes more vulnerable after it succeeds, not only before. Success brings institutions, institutions produce professions and interests, and interests extend the jurisdiction of the idea.

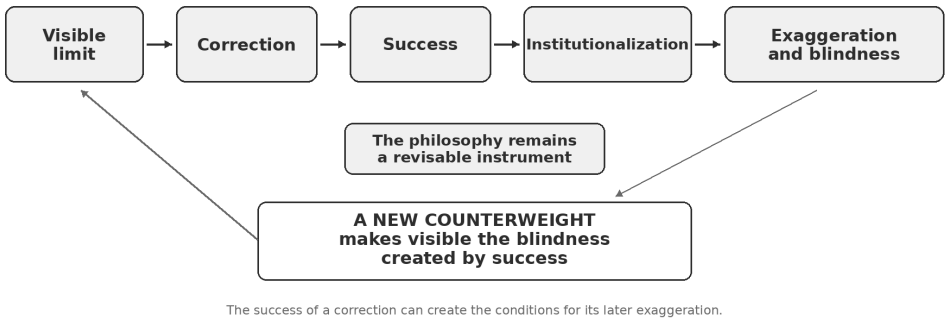

Diagram 7. The cycle of the counterweight

The diagram is not a law of history. It compresses a recurring dynamic. Ideas must be simplified in order to circulate and institutionalized in order to coordinate action. But simplification erases conditions, and the institution turns a perspective into roles, indicators, and authority. What began as liberation can become a new unrevisable normality.

Meta-rationality and outfinitism must be placed within this cycle, not above it. They arise in response to a culture in which rationalization, scaling, and automation have become so powerful that the limits of the frame are treated as noise. Meta-rationality corrects the identification of the model with reality. Outfinitism corrects the promise that every limit can be eliminated through more computation, more data, or a sufficiently large version of the same system.

Their value does not lie in replacing science, law, or engineering. They preserve the function of these institutions while withdrawing their claim to total jurisdiction. We preserve experiment and replication without pretending that measurement alone decides value. We preserve the predictability of law without pretending that procedure exhausts justice. We preserve optimization while making the objective and distribution of costs contestable. We preserve ideals while refusing to turn the impossibility of total conformity into an industry of concealment.

The two concepts can themselves become excessive. Meta-rationality can confer prestige on the person who speaks from several frames and marginalize simple action. Outfinitism can turn every limit into an excuse and every advance into a suspicious promise. Criticism of scientism can be captured by pseudoscience. Acceptance of

adaptive hypocrisy can protect cynicism. A culture of cones can become a world in which no one is accountable for claims because every contradiction is merely a change of perspective.

A philosophy serving as a counterweight must therefore contain an expiration clause. It is useful while it makes a relevant blindness visible and produces better decisions. It becomes suspect when it explains everything too easily, when its criticism can change nothing, when its vocabulary confers status, and when the cost of applying it disappears from the analysis. Meta-rationality must be sufficiently meta to see the moment when its additional level no longer helps. Outfinitism must be sufficiently finite to accept that its own role may come to an end.

This modesty does not weaken philosophy. It makes philosophy usable. A tool is not less valuable because it does not fit every problem. A hammer becomes dangerous when its success convinces us that everything is a nail. Meta-rationality and outfinitism do not offer the final form of wisdom. They may offer a necessary correction in an age when models produce social reality at scale and coherent content can be generated before the question has been maturely formulated.

A Culture of Transitions Between Cones in the Age of AI

Pluralistic societies cannot eliminate cones. A community needs shared language, procedures, identities, and reference points. A person cannot permanently examine every premise. An institution cannot renegotiate all its categories from the beginning before every decision. The problem is not the existence of lenses, but the capacity to move among them when reality demands it and to preserve the memory of the losses produced by the transition.

A culture of transitions would educate people to distinguish perspective from arbitrariness. It would show that a frame can be locally powerful without being universal. It would train the ability to formulate an opposing position without caricature and to state what evidence would change one's own view. It would not turn every lesson into an endless debate. It would preserve procedures in the domains where they work and make points of appeal visible.

In science, such a culture would communicate power and limits at the same time. The public would not be forced to choose between the myth of certainty and the cynicism that “experts are always changing their minds.” Revision would be presented as a normal function rather than a source of shame. Models would be tied to populations, scales, and conditions. Values involved in choosing thresholds and priorities would be separated from empirical results without being hidden behind them.

In law, the culture of transition would preserve the general rule while protecting judgment about the case. It would recognize that procedural equality can produce injustice when situations differ in relevant ways, yet avoid turning exceptions into opaque privilege. Contestation would be part of normal operation, not an attack on the institution. A rule that cannot be explained to those affected and admits no correction becomes a technique of power even if it was rationally designed.

In organizations, indicators would be treated as instruments with an expiration date. A metric would be accompanied by the question of which behaviors it may induce and what reality it fails to observe. When people learn to optimize the indicator, the institution would not merely accuse them of moral failure, but revise the frame. This does not eliminate individual responsibility. It distributes responsibility more realistically between the person and the incentive system.

AI makes this culture both necessary and possible. It is necessary because models can rapidly reproduce any cone sufficiently present in the data or constructed in a prompt. They can give every group a rhetorical machine that defends its world, generates examples without end, and adapts its message to the vulnerabilities of the interlocutor. Without literacy about frames, digital plurality does not automatically lead to openness; it can produce enclaves that are increasingly coherent and difficult to penetrate.

Yet AI can also reduce the cost of transition. A model can translate a technical argument into public language, show how the same problem appears to different actors, and separate claims about facts from claims about values. It can search for counterexamples, compare sources, and identify places where the same expression carries different meanings. It can help a person formulate an objection for which that person previously lacked the vocabulary.

Value appears only if the system preserves differences rather than dissolving them into artificial consensus. A good synthesis can say that two perspectives use the same data while assigning priority to different values. It can show where one of them contradicts evidence and where the conflict remains political. Sometimes translation leads to agreement. At other times it leads to clearer disagreement, which is progress as well.

In such a culture, alignment becomes a problem shared by humans and machines. We do not ask only how to make AI respect our values, as though those values were a stable file. We ask how to preserve the capacity to see when values conflict, when a frame becomes dominant through infrastructure, and when an optimization changes the society it was supposed to serve. A model cannot solve this problem for us, but it can be designed not to conceal it.

An AI oriented toward meta-rationality and outfinitism could contribute to this practice through small, verifiable gestures. It would make relevant ambiguity explicit, not every ambiguity. It would state when an answer is a hypothesis and when it is supported. It would seek tools before fabricating certainty. It would distinguish a person's experience from a causal explanation. It would adjust verification effort to the stakes. It would preserve alternatives and avenues of appeal where error can affect rights. It would know that more text does not mean more value.

These gestures can be imitated superficially. Institutions must therefore be aligned together with models. A product optimized exclusively for engagement does not become meta-rational through a few sentences of caution. An administration that does not permit contestation does not become outfinitist because it publishes the model's limitations. A company that concentrates authority does not solve the problem with a transparency dashboard. The technical cone is carried by a structure of power.

A culture of transitions therefore requires the separation of capability from authority. AI can be highly capable of generating, comparing, and predicting without receiving the right to decide alone. Authority should be granted gradually, by domain, with monitoring, alternatives, and the possibility of withdrawal. This is not fear of technology. It is the normal way in which a mature society treats any new power.

It also requires the preservation of human capacities. Automation can eliminate repetitive work and expand access to expertise. It can also atrophy the judgment required for contestation. A person who is formally “in the loop” does not control the system if that person no longer has the time, knowledge, or institutional support needed to say no. Education must cultivate precisely the capacities that AI makes more important: framing the problem, evaluating sources, recognizing the frame, making value judgments, and assuming responsibility for consequences.

Meta-rationality and outfinitism can become a language for these capacities. They are neither the only possible language nor the first. Fallibilism, pragmatism, systems theory, practical wisdom, and social epistemology have articulated important parts of the problem. Their possible contribution lies in joining movement among frames with the discipline of real frontiers at a moment when AI makes both problems impossible to ignore.

Wisdom does not mean a final escape from every cone. Such a position would be a cone denying its own shape. Wisdom means inhabiting a frame long enough to act while preserving enough distance to notice when reality exceeds it. Philosophies are useful when they make that distance possible. They become dangerous when they demand to be inhabited forever.

CONCLUSION

SEEING THE WORLD AND SEEING THE CONE

This book began with a contemporary experience and an old intuition. The experience is that a language model can receive an incompletely formulated philosophy and produce developments that appear to preserve its meaning. The intuition is that beyond the details of a theory there is sometimes a way of seeing that travels across domains, epochs, and institutions. Relativity offered culture a grammar of perspective. Evolution made selection and adaptation available as general explanations. Cybernetics spread feedback. Rationality became science, law, administration, and a moral ideal. In every case, transfer produced both discovery and deformation.

The metaphor of the cone of meaning attempted to compress this dynamic. A cone does not reproduce the world. It selects a region of proximity and makes certain continuations more natural. In a language model, context can orient representations and probabilities toward a family of concepts and styles. In a culture, institutions, language, and memory make some explanations more visible. In a person, role, identity, and purpose change what matters. The cone is not a unique entity hidden in the brain or in history. It is an image of the selective orientation without which thought would be impossible.

This selectivity is the source of both intelligence and blindness. A model is useful because it leaves things aside. A procedure is efficient because it reduces variety. An ideal coordinates because it simplifies. Error appears when we forget the compression and grant the instrument total jurisdiction. Scientism confuses the power of science with the capacity to decide every value. Legalism confuses procedure with complete justice. Optimization confuses the indicator with the good. Moral idealism confuses direction with the possibility of perfect conformity and produces adaptive hypocrisy.

Meta-rationality and outfinitism have been proposed as two counterweights. The first demands rigor while including purpose, scale, time, power, and alternatives within the examination. It remains situated, without claiming an absolute perspective. The second brings this movement into the world of resources and consequences: every real act is finite, every validation has a frontier, and every institution has limited capacity. The horizon can nevertheless remain open. A declared limit is not capitulation, but the starting point for controlled extension.

They complement one another like two movements of breathing. Meta-rationality opens the field and prevents premature closure. Outfinitism stops provisionally, ties the claim to its coverage, and permits a decision. Without openness, discipline can serve the wrong frame. Without limits, plurality can become regress, status, or paralysis. Together they do not produce the final system, but a more mature practice of moving among systems.

The LLM makes this practice urgent. It shows that meaning has enough relational structure for an orientation to be reconstructed from clues, contrasts, and purposes. The prompt can function as a local constitution and make one cone more probable. The same mechanism enables translation, creation, and intellectual assistance. It also enables the rapid manufacture of coherence, the amplification of a false premise, and sycophancy. The fact that the model gives us the feeling that it has understood is both a functional achievement and a source of risk. It should not be reduced to "mere statistics," but neither should it be turned into evidence of a digital soul.

Abundant generation shifts scarcity. It is no longer enough to be able to produce an idea, a report, a program, or an image. We must decide what deserves to be verified, preserved, and put into action. Epistemology and axiology become inseparable. Truth does not choose the goal by itself, and value cannot ignore truth. An AI can generate what is plausible; society must preserve the difference among the plausible, the justified, the legitimate, and the good.

From this follows the hypothesis of alignment as a state of orientation. The phrase "state of consciousness" remains a risky metaphor and should be used only to suggest a global organization of relevance, not demonstrated phenomenal experience. The defensible form of the claim says that an AI is more than the sum of the prohibitions it obeys. Its behavior depends on how it interprets the situation, distributes attention, handles the unknown, and orders values. Post-training, the prompt, the tools, and the institution shape its orientation.

A meta-rational and outfinitist filter could improve this orientation. It could reduce unjustified certainty, confusion between analogy and mechanism, sycophancy, the expansion of authority, and the concealment of limits. It could encourage the model to distinguish facts from value choices, adjust verification to the stakes, and preserve a path of correction. It would not solve alignment. It can be imitated, poorly measured, captured by the provider, or transformed into ritual refusals. Many risks arise from the organization that uses the system, not from the model's words.

The traps are not a pessimistic appendix, but a constitutive part of the thesis. The cone can become apophenia. Meta-rationality can become relativism or elitism. Outfinitism can become an alibi. Criticism of scientism can nourish anti-science.

Understanding hypocrisy can justify cynicism. A plurality of explanations can forget the victim. An institution can speak about limits without redistributing authority. A philosophy that does not keep these risks visible contradicts its own purpose.

The mature form of the idea does not ask people to abandon convictions. It asks convictions to state where they are strong, what they have compressed, and what would change them. It does not ask society to abandon science, law, or rationality. It asks that their success not be turned into a claim to totality. It does not ask AI to see without a filter. It asks that filters become more inspectable, be changeable when the stakes require it, and not receive authority from fluency alone.

Every new philosophy is valuable chiefly as a rebalancing. It observes an exaggeration of the old lens, recovers a forgotten dimension, and creates a vocabulary for correction. If it succeeds, it will in turn be simplified, institutionalized, and exaggerated. This is not an accidental anomaly. Ideas must take form in order to circulate, and form produces inertia. Wisdom does not consist in inventing the frame immune to this cycle, but in building a culture that recognizes the cycle earlier.

We cannot live outside representations. Reality is not available to us without selection, language, and purpose. We can, however, avoid confusing the map with the territory, local success with total authority, and fluency with wisdom. We can use ideals without sacralizing appearances and technology without surrendering our purposes to it.

Wisdom does not stand above the cones. It is the freedom to see their cost without losing the power to build through them. We inevitably see through cones; sometimes also seeing the cone may be the beginning of a more mature reason.

REFERENCES

[ALBOAIE-2026] Alboaie, Sînică. Outfinitism: Meta-Rationality and the Limits of Every Frame. Toward a Meta-Science of Limits. Consolidated manuscript, 2026.

[ALBOAIE-2026-MATH] Alboaie, Sînică. Outfinite Mathematics for Executable Science. Research position book, version 0.5, August 2026.

[ASHBY-1956] Ashby, W. Ross. An Introduction to Cybernetics. Chapman & Hall, 1956.

[BAI-2022] Bai, Yuntao et al. "Constitutional AI: Harmlessness from AI Feedback." arXiv:2212.08073, 2022. https://arxiv.org/abs/2212.08073

[BALTES-STAUDINGER-2000] Baltes, Paul B., and Ursula M. Staudinger. "Wisdom: A Metaheuristic (Pragmatic) to Orchestrate Mind and Virtue toward Excellence." American Psychologist 55, no. 1, 2000, pp. 122–136.

[BENDER-KOLLER-2020] Bender, Emily M., and Alexander Koller. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." Proceedings of ACL 2020, pp. 5185–5198. https://aclanthology.org/2020.acl-main.463/

[BETLEY-2025] Betley, Jan et al. "Emergent Misalignment: Narrow Finetuning Can Produce Broadly Misaligned LLMs." arXiv:2502.17424, 2025. https://arxiv.org/abs/2502.17424

[BRANG-2025] Brang, Michael, Franziska Greinert, Malte S. Ubben, Helena Franke, and Philipp Bitzenbauer. "Quantum Terminology in Pseudoscience: A Review and Analysis of the Use of Quantum Physics Concepts in Alternative Medicine." EPJ Quantum Technology 12, 96, 2025. https://doi.org/10.1140/epjqt/s40507-025-00403-9

[BUSKELL-2017] Buskell, Andrew. "What Are Cultural Attractors?" Biology & Philosophy 32, 2017, pp. 377–394. https://doi.org/10.1007/s10539-017-9570-6

[CARTWRIGHT-1983] Cartwright, Nancy. How the Laws of Physics Lie. Oxford University Press, 1983.

[CLAIDIERE-SPERBER-2007] Claidière, Nicolas, and Dan Sperber. "The Role of Attraction in Cultural Evolution." Journal of Cognition and Culture 7, nos. 1–2, 2007, pp. 89–111.

[DAMASIO-1994] Damasio, Antonio R. Descartes' Error: Emotion, Reason, and the Human Brain. Putnam, 1994.

[DASTON-GALISON-2007] Daston, Lorraine, and Peter Galison. Objectivity. Zone Books, 2007.

[DAWKINS-1976] Dawkins, Richard. The Selfish Gene. Oxford University Press, 1976.

[DEVLIN-2019] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Proceedings of NAACL-HLT 2019. https://aclanthology.org/N19-1423/

[ETHAYARAJH-2019] Ethayarajh, Kawin. "How Contextual Are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings." Proceedings of EMNLP-IJCNLP 2019, pp. 55–65. https://aclanthology.org/D19-1006/

[EU-2024] European Union. Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence. Official Journal of the European Union, 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[FLECK-1935] Fleck, Ludwik. Genesis and Development of a Scientific Fact. University of Chicago Press, 1979; original edition, 1935.

[FOUCAULT-1966] Foucault, Michel. The Order of Things: An Archaeology of the Human Sciences. Pantheon Books, 1970; original edition, 1966.

[GALISON-2003] Galison, Peter. Einstein's Clocks, Poincaré's Maps: Empires of Time. W. W. Norton, 2003.

[GIERE-2006] Giere, Ronald N. Scientific Perspectivism. University of Chicago Press, 2006.

[GOODHART-1975] Goodhart, Charles A. E. "Problems of Monetary Management: The U.K. Experience." In Papers in Monetary Economics, vol. 1. Reserve Bank of Australia, 1975.

[GUAN-2024] Guan, Melody Y. et al. "Deliberative Alignment: Reasoning Enables Safer Language Models." arXiv:2412.16339, 2024. https://arxiv.org/abs/2412.16339

[HACKING-1992] Hacking, Ian. "Style for Historians and Philosophers." Studies in History and Philosophy of Science Part A 23, no. 1, 1992, pp. 1–20.

[HENDEL-2023] Hendel, Roee, Mor Geva, and Amir Globerson. "In-Context Learning Creates Task Vectors." Findings of EMNLP 2023, pp. 9318–9333. https://aclanthology.org/2023.findings-emnlp.624/

[KING-KITCHENER-1994] King, Patricia M., and Karen Strohm Kitchener. Developing Reflective Judgment. Jossey-Bass, 1994.

[KING-KITCHENER-2004] King, Patricia M., and Karen Strohm Kitchener. "Reflective Judgment: Theory and Research on the Development of Epistemic Assumptions through Adulthood." Educational Psychologist 39, no. 1, 2004, pp. 5–18.

[KNORR-CETINA-1999] Knorr Cetina, Karin. Epistemic Cultures: How the Sciences Make Knowledge. Harvard University Press, 1999.

[KUHN-1962] Kuhn, Thomas S. The Structure of Scientific Revolutions. University of Chicago Press, 1962.

[LI-2022] Li, Kenneth et al. "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task." arXiv:2210.13382, 2022. https://arxiv.org/abs/2210.13382

[LIGHTMAN-2023] Lightman, Hunter et al. "Let's Verify Step by Step." arXiv:2305.20050, 2023. https://arxiv.org/abs/2305.20050

[LONGINO-1990] Longino, Helen E. Science as Social Knowledge: Values and Objectivity in Scientific Inquiry. Princeton University Press, 1990.

[MITCHELL-2009] Mitchell, Sandra D. Unsimple Truths: Science, Complexity, and Policy. University of Chicago Press, 2009.

[NIST-2023] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1, 2023. https://doi.org/10.6028/NIST.AI.100-1

[OUYANG-2022] Ouyang, Long et al. "Training Language Models to Follow Instructions with Human Feedback." Advances in Neural Information Processing Systems 35, 2022. https://arxiv.org/abs/2203.02155

[PAN-2023] Pan, Jane, Tianyu Gao, Howard Chen, and Danqi Chen. "What In-Context Learning 'Learns' In-Context: Disentangling Task Recognition and Task Learning." Findings of ACL 2023, pp. 8298–8319. https://aclanthology.org/2023.findings-acl.527/

[PENNINGTON-2014] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation." Proceedings of EMNLP 2014, pp. 1532–1543. https://aclanthology.org/D14-1162/

[SHARMA-2023] Sharma, Mrinank et al. "Towards Understanding Sycophancy in Language Models." arXiv:2310.13548, 2023. https://arxiv.org/abs/2310.13548

[SIMON-1955] Simon, Herbert A. "A Behavioral Model of Rational Choice." Quarterly Journal of Economics 69, no. 1, 1955, pp. 99–118.

[SPERBER-1996] Sperber, Dan. Explaining Culture: A Naturalistic Approach. Blackwell, 1996.

[VASWANI-2017] Vaswani, Ashish et al. "Attention Is All You Need." Advances in Neural Information Processing Systems 30, 2017. https://arxiv.org/abs/1706.03762

[WANG-2025] Wang, Miles et al. "Toward Understanding and Preventing Misalignment Generalization." OpenAI, 2025. https://openai.com/index/emergent-misalignment/

[WEBER-1922] Weber, Max. Economy and Society. University of California Press, 1978; original edition, 1922.

[WHITWORTH-2001] Whitworth, Michael H. Einstein's Wake: Relativity, Metaphor, and Modernist Literature. Oxford University Press, 2001.