universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Do Artificial Neural Networks Know? A Philosophical Inquiry into Their Epistemic Status with Particular Attention to Propositional and Tacit Knowledge

verfasst von | submitted by

Lionel Augustin Thalmann BA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2026

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 941
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Philosophie
Betreut von | Supervisor:	Dr. Michael Funk B.A. M.A.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, *Dr. Michael Funk*, for his consistently competent guidance and thoughtful support throughout the entire research and writing process. Dr. Funk is among the few scholars whose expertise spans both philosophy and computer science, and I have benefited greatly from his ability to connect conceptual questions with technical detail, as well as from his careful feedback at every stage of this project.

I am also grateful to *Prof. Dr. Peter Reichl* and to the *Cooperative Systems (COSY)* Research Group at the Institute of Computer Science at the University of Vienna for providing an intellectually stimulating environment in which research can genuinely extend beyond disciplinary boundaries. The group's openness to diverse perspectives and its commitment to rigorous, collaborative inquiry offered an important context for developing the ideas pursued in this thesis.

My heartfelt thanks also go to my dear brother, *Philemon Thalmann, MSc*, for his competent guidance on the mathematical aspects of this project, particularly in areas related to the theory of machine learning, and to my dear friend, *Matteo Albrecht, MA,* for his philosophical insights and for carefully reviewing the manuscript.

My heartfelt thanks also go to my dear mother, *Barbara Weber*, and my dear father, *Michael Thalmann*, for carefully reading drafts of this work and for their constructive comments.

Finally, and most importantly, I am deeply grateful to my beloved wife, *Aleksandra Ilyina*, for her patience, encouragement, and unwavering support throughout this process.

Abstract

Contemporary artificial neural networks deliver fluent classifications, explanations, and recommendations that increasingly function as epistemic intermediaries. This thesis asks whether—and in what sense—a subsymbolic learner (CLERE) can be said to know. Methodologically, it pursues an epistemological pluralism that integrates a technical reconstruction of CLERE's competence (error-driven learning, distributed representations, and high-dimensional encoding) with two complementary epistemic vocabularies: propositional knowledge (the justified-true-belief schema and post-Gettier debates over internalism and externalism) and tacit knowledge (Ryle's knowing-how, Polanyi's tacit knowing, and Collins's taxonomy of relational, somatic, and collective tacit knowledge).

The assessment yields a bifurcated verdict. On the propositional side, classical JTB does not straightforwardly apply, since belief and justification are traditionally tied to a doxastic perspective in the space of reasons; nonetheless, knowledge-like attributions can be reconstructed either internally, as calibrated credences supported by local evidential checks, or externally, by treating CLERE as a belief-forming process whose outputs are warranted only when embedded in robust socio-technical reliability indicators and scientific anchoring. On the tacit side, CLERE plausibly exhibits a thin, performance-based know-how within its trained domain, yet it fails to satisfy thicker Polanyian tacit knowing: its integrations are mathematically optimized rather than intentionally meant, it lacks embodied indwelling and vulnerability, and opacity alone is not tacitness. Finally, Collins's socialization problem marks a residual limit: continual learning from curated data traces does not amount to normative participation and answerability within a living practice.

The upshot is that CLERE is best understood as an epistemic instrument whose outputs can be knowledge-bearing within appropriately governed pipelines, while the primary bearer of epistemic competence is the socio-technical practice that trains, validates, and corrects the system.

Abstract (German)

Moderne künstliche neuronale Netze erzeugen Klassifikationen, Erklärungen und Empfehlungen und übernehmen damit zunehmend die Rolle epistemischer Vermittlungsinstanzen. Die vorliegende Arbeit untersucht, ob—und in welchem Sinne—einem subsymbolischen Lerner (CLERE) Wissen zugeschrieben werden kann. Methodisch folgt sie einem epistemologischen Pluralismus, der eine technische Rekonstruktion der Kompetenz von CLERE—fehlergetriebenes Lernen, verteilte Repräsentationen und hochdimensionale Kodierung—mit zwei komplementären epistemischen Vokabularen zusammenführt: propositionalem Wissen (dem JTB-Schema sowie den Post-Gettier-Debatten über Internalismus und Externalismus) und implizitem Wissen (Ryles *Wissen-wie*, Polanyis *implizites Wissen* und Collins' Taxonomie des *relationalen*, *somatischen* und *kollektiven* impliziten Wissens).

Die Analyse führt zu einem zweigeteilten Ergebnis. Auf der propositionalen Seite lässt sich das klassische JTB-Schema nicht unmittelbar anwenden, da Überzeugung und Rechtfertigung traditionell eine doxastische Perspektive im *Raum der Gründe* (space of reasons) voraussetzen. Wissensanaloge Zuschreibungen lassen sich dennoch rekonstruieren: internalistisch als kalibrierte Überzeugungsgrade (*credences*), gestützt auf lokale Evidenzkontrollen; externalistisch, indem CLEREs Ausgaben als Produkte eines Prozesses der Überzeugungsbildung betrachtet werden, deren epistemische Berechtigung von ihrer Einbettung in robuste soziotechnische Zuverlässigkeitsindikatoren und eine wissenschaftliche Fundierung abhängt. Auf der Seite des impliziten Wissens zeigt CLERE zwar ein performanzbasiertes Können im schwachen Sinne innerhalb seiner trainierten Domäne, erfüllt jedoch nicht die anspruchsvolleren Kriterien des polanyischen Begriffs von *implizitem Wissen*: Seine Integrationsleistungen sind mathematisch optimiert, nicht intentional gemeint; es mangelt ihm an Einfühlung (indwelling) und Verletzbarkeit; Intransparenz allein konstituiert noch kein implizites Wissen. Collins' Sozialisationsproblem markiert schließlich eine verbleibende Grenze: Kontinuierliches Lernen aus kuratierten Datenspuren ersetzt keine normative Teilhabe an einer lebendigen Praxis mit ihren Rechenschaftspflichten.

Das Ergebnis lautet: CLERE ist am besten als epistemisches Instrument zu verstehen, dessen Ausgaben innerhalb angemessen geregelter Verarbeitungsketten wissenstragend sein können—während der primäre Träger epistemischer Kompetenz die soziotechnische Praxis bleibt, die das System trainiert, validiert und korrigiert.

Contents

Notation

This is a quick reference for frequently used abbreviations and mathematical symbols. Infrequent abbreviations are defined where they first appear.

Abbreviations

Artificial Intelligence & Machine Learning

AI	Artificial Intelligence
ANN	Artificial Neural Network
DL	Deep Learning
GPT	Generative Pre-trained Transformer
LLM	Large Language Model
LSTM	Long Short-Term Memory (network)
MLP	Multilayer Perceptron
NLP	Natural Language Processing
RNN	Recurrent Neural Network
SLP	Single-Layer Perceptron
UAT	Universal Approximation Theorem

Case Study

CLERE	Conceptual Learner for Epistemic Reasoning Experiments

Epistemology & Philosophy

JTB	Justified True Belief
PR	Process Reliabilism
PSR	Principle of Sufficient Reason
TK	Tacit Knowledge
RTK	Relational Tacit Knowledge

STK	Somatic Tacit Knowledge
CTK	Collective Tacit Knowledge
SLTK	Somatic-Limit Tacit Knowledge
SATK	Somatic-Affordance Tacit Knowledge

Tokens, Labels, and Metrics

BOS	Beginning-of-sequence token (e.g., `<BOS>`, `<BOS-RU>`)
EOS	End-of-sequence token (e.g., `<EOS>`)
BCE	Binary Cross-Entropy
XOR	Exclusive OR
GPS	General Problem Solver
LT	Logic Theorist

Mathematical Notation

In our reconstruction of neural network architectures (Chapter 2), we follow the mathematical notation and didactic exposition of Hofmann (2022) as the primary reference for threshold functions, perceptrons, multilayer networks, backpropagation, recurrent architectures, and transformer projections; supplemented by Shalev-Shwartz and Ben-David (2014) for general machine-learning theory and by Vaswani et al. (2017) for scaled dot-product self-attention. Figures for the attention mechanism and attention-score variants follow Voita (2020); the RNN and LSTM diagrams are adapted from Starmer (2025).

Sets and Indices

R	Set of real numbers
Z	Set of integers
t	Time-step index (recurrent models)
i, j	Sequence-position indices

Neural Networks and Optimization

x	Input (vector)
x_t	Input at time step t
y	Target label
$\hat{y}$	Predicted label
θ	Model parameters
W, b	Weight matrix and bias vector
$\phi(\cdot)$	(Generic) nonlinear activation function
$\sigma(\cdot)$	Logistic sigmoid function
$\mathrm{ReLU}(\cdot)$	Rectified linear unit activation
h_t	Hidden state at time step t
C_t	LSTM cell state at time step t
$\mathcal{L}$	Loss function
η	Learning rate
p	Model output probability (binary) or probability vector over classes
K	Number of classes (multiclass setting), where context makes this usage explicit

Embeddings, Sequences, and Attention

$\mathcal{V}$	Vocabulary (set of tokens)
$V = \lvert\mathcal{V}\rvert$	Vocabulary size (scalar)
n	Sequence length
d	Embedding / hidden-state dimensionality
$X \in R^{n \times d}$	Matrix of token embeddings for a length-n sequence
W^Q, W^K, W^V	Query, key, and value projection matrices
Q, K, V	Query, key, and value matrices (in attention; V is a matrix here)
S	Attention score matrix ($S = QK^\top$)
d_k	Query/key dimensionality (in scaled dot-product attention)
$\mathrm{softmax}(\cdot)$	Row-wise normalization to a probability distribution
$\mathrm{Attention}(Q, K, V)$	Scaled dot-product attention

Probability and Information-Theoretic Quantities

$P(A)$	Probability / credence assigned to proposition or event A
$P(A \mid B)$	Conditional probability of A given B
$S(p)$	Surprisal of an event with probability p
$K \subseteq R^n$	Compact set (in the statement of the UAT), where context makes this usage explicit

1. Introduction

When Gulliver, from Jonathan Swift's novel *Gulliver's Travels* (1726), is brought before the giant king of Brobdingnag, a leader well educated in both philosophy and mathematics, the king suspects—judging by Gulliver's strange size relative to his own—that he might be some sort of cleverly orchestrated machine: "[...] a piece of clock-work (which is in that country arrived to a very great perfection) contrived by some ingenious artist" (Swift 2005, p. 124). Even after the king hears Gulliver speak—and Gulliver can just about manage to converse in their language—the king is by no means convinced that Gulliver is in fact a real human being. The king suspects that the engineer who created Gulliver has taught him a set of words to make him "sell at a better price," and therefore commissions his greatest scholars to test Gulliver with their questions, challenging his ability to reason. Only after they receive nothing but "rational answers" from Gulliver do they conclude unanimously that Gulliver is in fact not a machine—just some sort of *lusus naturae* ("freak of nature").[1]

In the context of the questions that we are about to pursue, there are two interesting aspects to this excerpt of Swift's novel: First, we can see that the conceptual ambiguity surrounding how to properly denote Gulliver—whether as a machine and therefore as something that could be bought and sold as a cleverly engineered commodity or as a human being deserving of recognition as a rational agent—is not merely a contemporary phenomenon but, although confined to the realm of fiction, already present as a philosophical puzzle in 18th-century literature. Even though the technology of Swift's time period offered no plausible route to language-producing artifacts, the uncertainty about how to categorize an entity that exhibits rational linguistic behavior nonetheless functioned as a provocative limit case for reflection on what, if anything, separates merely mechanical behavior from genuinely rational discourse.

Second, we notice that Swift's figure of the king suggests a specific method by which to disambiguate this conceptual uncertainty: he proposes to test the capacity to sustain appropriate linguistic exchange as the decisive criterion. This idea—that conversational competence might serve as a test for distinguishing "artificial" from "genuine" intelligence—recurs throughout modern philosophy and literature, and it eventually

[1] "A determination [...]", so Swift remarks, "[...] exactly agreeable to the modern philosophy of Europe, whose professors, disdaining the old evasion of occult causes, whereby the followers of Aristotle endeavoured in vain to disguise their ignorance, have invented this wonderful solution of all difficulties, to the unspeakable advancement of human knowledge" (Swift 2005, pp. 125–126).

becomes explicit as a methodological proposal in twentieth-century discussions of artificial intelligence. Most influentially, Turing's *Imitation Game* proposes to set aside the theoretically loaded question "Can machines think?" and to replace it with an operational challenge: can a machine's linguistic performance be such that a competent human judge, limited to text-based interaction, cannot reliably tell it apart from a human interlocutor (Turing 1950)?[2]

To better understand the stakes of this question, let us take a closer look at what the term "machine" signifies in these different contexts. A machine in the time of Swift was fundamentally a mechanical apparatus—an assemblage of gears, levers, springs, and "clockwork" mechanisms whose behavior, however intricate, remained reducible to the deterministic principles of classical mechanics. What Turing had in mind, by contrast, were early versions of what is known today as *symbolic* or *rule-based* AI: "machines" that worked by manipulating formal symbols according to predetermined logical rules—an approach that achieved notable successes compared to its mechanical predecessors, yet also revealed profound limitations. Such systems tended to be brittle outside carefully specified contexts; they struggled with the background competence that humans bring to ordinary conversation; and they made painfully visible how much of linguistic understanding depends on context, pragmatics, and shared world knowledge that is not easily reduced to explicit, enumerated premises.

The "intelligent machines" we have today—and that will be the central subject of this thesis—represent a fundamental departure from both of these conceptions. Contemporary artificial intelligence systems, particularly those based on deep neural networks, have emerged through a dramatic shift from rule-based symbol manipulation to *machine learning*: statistical models trained on vast corpora of data that develop their own internal representations through processes of iterative optimization. These systems—specifically, large-scale artificial neural networks—are capable of producing "rational answers" on demand in ways that would satisfy both the most exacting of the giant king's scholars as well as Turing's imitation test. They can explain, argue, diagnose, advise, and teach with an ease that rivals highly educated human experts across an expanding range of domains.

[2] The point of Turing was not so much that successful imitation would settle the philosophical question, but rather that it would make the dispute answerable in practice: instead of debating what "thinking" must be, he asks what conversational abilities would count as an adequate stand-in for it under controlled conditions (Turing 1950).

Machine Epistemology

The trajectory from Swift's clockwork to Turing's symbol manipulators to contemporary systems capable of producing human-like answers, sustaining linguistic exchanges, and exhibiting context-sensitive intelligent behavior does not merely represent a historical shift in what a "machine" is taken to be; it marks a shift in the epistemic roles that such systems can occupy as well. These artifacts do not simply perform tasks *in* epistemic environments; they increasingly function *as* epistemic intermediaries, shaping what information is available, how it is framed, and what users come to believe. This brings into focus questions that standard epistemological templates were not built to handle: Where, if anywhere, is the relevant epistemic subject to be located—in the trained model, in its deployment context, or in the larger sociotechnical pipeline that selects data, defines objectives, and validates performance? What could justification amount to when a system's "warrants" are distributed across high-dimensional representations that resist transparent articulation? How should we assess reliability when outputs are probabilistic, sensitive to prompts and context, and vulnerable to distribution shift and confabulation? And how should epistemic trust, authority, and responsibility be calibrated when fluent answers can function as a form of quasi-testimony without the familiar normative capacities of a human knower?

This cluster of problems motivates what has come to be known in contemporary philosophy as *machine epistemology*—the systematic investigation of knowledge and epistemic relations in the context of artificial intelligence (Gramelsberger and Sigwart 2025, p. 261). As artificial neural networks achieve remarkable performance on tasks traditionally associated with human intelligence—from language understanding to mathematical reasoning—they challenge longstanding philosophical assumptions about the relationship between behavioral competence and epistemic capacity. The success of these systems raises a fundamental puzzle: can we meaningfully ascribe knowledge to systems whose internal mechanisms differ radically from human cognitive architecture, whose "understanding" may consist in statistical patterns across high-dimensional vector spaces rather than conscious grasp of concepts, and whose reliability depends on training conditions and deployment contexts in ways that complicate traditional notions of epistemic justification?

Our specific focus within this emerging field is the question of whether and under what conditions artificial neural networks can meaningfully be said to possess *knowledge*. We aim to develop a rigorous epistemological assessment that is grounded in both the technical realities of how ANNs function and the conceptual sophistication required to avoid mistaking surface-level performance for genuine epistemic achievement.

1.1. Terminological Preliminaries

The terms employed in discussions about AI—terms like "knowledge," "intelligence," "understanding," "learning," and indeed "AI" itself—are used with remarkably different meanings across (and even within) academic and non-academic disciplines. In technical contexts—in machine learning research, in industry applications, and in the development of commercial AI products—a machine learning system is often said to "know" a domain when it has been trained on relevant data; it "understands" natural language when it can process and generate linguistically appropriate responses; and it possesses "intelligence" when it performs well on benchmark tasks. In philosophical discourse, by contrast, these terms are typically situated within long conceptual traditions, sharpened by distinctions and assessed by standards that are not reducible to benchmark performance—but, then again, this conceptual refinement often comes at the price of proceeding without a working grasp of the technical and mathematical concepts that determine what an "AI" system actually is.

These terminological imprecisions create situations where substantive disagreements are often difficult to distinguish from merely verbal disputes, where empirical claims are conflated with conceptual ones, and where philosophical analysis proceeds without sufficiently clear definitions of its subject matter. What we aim to accomplish in this thesis is therefore to develop an analytical framework that is adequate both to the technical realities of how machine learning systems actually function and to the philosophical sophistication required to avoid mistaking surface-level performance for genuine epistemic achievement when we apply terms like "intelligence" or "knowledge" to such systems. In order to do so, we begin by briefly clarifying the key terms and the distinctions between them that will structure our subsequent analysis.

The first term most directly implicated in the phrase "artificial intelligence" is *intelligence* itself. The term "intelligence" admits of multiple interpretations, but in the context of AI, it is typically understood as a general capacity for problem-solving, adaptation, and goal-directed behavior (Legg and Hutter 2007). Intelligence, on this view, is something like a measure of an agent's effectiveness in achieving its objectives across diverse environments. While intelligence clearly involves cognitive capacities, it does not necessarily entail knowledge in the epistemic sense we will be articulating. An intelligent system might excel at optimization tasks, pattern recognition, or strategic planning without possessing knowledge in any robust sense. Indeed, much of what we call "artificial intelligence" consists of systems that exhibit impressive problem-solving capabilities while operating on

purely statistical correlations that may or may not correspond to genuine understanding of underlying causal structures or semantic content.[3]

The relationship between intelligence and knowledge is thus neither one of identity nor of simple entailment. One might possess extensive knowledge without being particularly intelligent (consider someone with vast memorized information but poor reasoning skills), and one might exhibit intelligent behavior without possessing knowledge in any demanding epistemological sense. A chess engine that evaluates millions of positions per second and consistently defeats human grandmasters exhibits remarkable intelligence in the domain of chess, but whether it "knows" anything about chess—whether it grasps chess concepts, understands strategic principles, or possesses justified beliefs about good moves—is yet to be determined.

One concept commonly invoked to bridge the gap between mere intelligent behavior and genuine epistemic achievement is *understanding*. Understanding has emerged as a central concept in recent epistemological literature (Grimm 2021; Elgin 2017). While knowledge is often analyzed in terms of content (knowing *that* p or knowing *how* to ϕ), understanding involves grasping how pieces of information hang together in a way that makes a phenomenon intelligible—paradigmatically, grasping *why* things are as they are. One influential strategy is to define understanding in terms of the ability to manipulate representations or engage in counterfactual reasoning: to understand a domain is to be able to answer "what if" questions about it, to see how changes in conditions would alter outcomes, and to extend what one knows to novel cases in a principled way (Kelp 2015). On such views, understanding is not exhausted by possessing isolated true propositions; it is closer to having a structured, explanatory "grip" on a subject matter that supports inference, prediction, and evaluation. Accordingly, the relationship between knowledge and understanding is complex and contested. Some argue that understanding is simply a species of knowledge—specifically, knowledge of explanations or knowledge of how elements relate to one another (Grimm 2011; Kelp 2015). Others maintain that understanding is a distinct epistemic state that, while related to knowledge, cannot be reduced to it (Elgin 2017).

By moving from the more general term "AI" toward the more concrete field of *machine learning*—or *deep learning*—another important concept is *learning*, which in the machine

[3]It is for this reason that several authors have suggested replacing or supplementing the term "intelligence" with alternative characterizations such as "information processing" or "computational pattern recognition" that make fewer commitments to cognitive or epistemic capacities (Funk 2023; Bostrom 2014; Chollet 2019).

learning literature refers to the process by which a system improves its performance on a task through experience. Learning is clearly relevant to knowledge—it is one of the primary means by which knowledge is acquired. But learning and knowledge are distinct concepts. One can learn (in the sense of adjusting one's behavior based on feedback) without acquiring knowledge, as when a thermostat "learns" to regulate temperature without knowing anything about thermodynamics. Conversely, one might possess knowledge without having learned it in any ordinary sense—as when one knows certain logical truths or a priori propositions simply by grasping them.

Finally, the term "knowledge" itself, as we shall see, is contested and internally differentiated. It is not a monolithic concept but encompasses several distinct forms. For the purposes of this thesis, we will focus on two fundamental categories within this field: *propositional knowledge* and *tacit knowledge*. Propositional knowledge (knowledge-that) is knowledge expressible in declarative sentences—knowing that Paris is the capital of France, that water is H_2O, that $2 + 2 = 4$—and tacit knowledge is knowledge embedded in skills and practices that cannot be fully articulated in propositional form—the kind of knowledge, for example, that enables skilled action without being able to fully explicate the principles involved.

1.2. Methodology

The central methodological commitment underlying this thesis is that epistemology is, in its essence, plural (Gramelsberger and Sigwart 2025, p. 14). There is no single, privileged framework that exhausts what it means to know, no via unica that captures all the ways in which cognitive subjects—whether biological or artificial—can stand in appropriate epistemic relations to truth. Different epistemological frameworks each illuminate different aspects of epistemic normativity, and each may be more or less adequate to different contexts of evaluation. With this conviction, we touch upon a broader methodological reorientation in the study of knowledge that has developed over the past decades.

Classical epistemology has traditionally taken its primary orientation from propositional, linguistically articulated knowledge—*knowing-that*—and has consequently focused on questions of justification, truth, and belief within that register. As Abel and Conant 2012 and others have argued, this orientation renders classical epistemology incapable of

adequately accommodating other modes of knowing that are no less fundamental: procedural and practical knowledge (*knowing-how*), non-conceptual knowledge (perceptual, intuitive, emotional), and implicit or tacit knowledge that operates as a precondition of explicit cognition without itself being fully articulable in propositional form. In light of what Abel diagnoses as the "dominant cognitivist bias" (ibid., p. 3) of contemporary epistemology—i.e., constraining the thinking about knowledge to only (this) one dimension of the epistemic—we take seriously his notion of *rethinking* the scope and methods of epistemological inquiry.

According to Abel, before we can meaningfully pose epistemological questions about justification, validity, and the limits of knowledge, we must first establish which *forms* of knowledge—and which modes of their interplay—are at issue in a given case. This task corresponds to what he denotes as *knowledge research* (ibid., p. 1): Knowledge research is concerned with describing the distinctive profiles of various modes of knowing, identifying their points of overlap, elucidating the mechanisms by which they interact and interpenetrate, and grasping the dynamic character of these interactions as they manifest in perception, language, thought, and action (ibid., p. 8).

Although what we are about to undertake with the following thesis is not knowledge research in Abel's full and ambitious sense—our focus lies not so much in investigating the mechanisms by which different knowledge forms interact but in developing a concrete taxonomy of epistemological frameworks (propositional knowledge and tacit knowledge) and applying this taxonomy to a specific epistemic object, namely a subsymbolic artificial neural network—Abel's program nonetheless provides the methodological horizon against which our approach becomes intelligible. Concretely, we consider three aspects of his analysis to be of particular importance for our purposes.

First, the irreducible plurality of knowledge forms. The two forms of knowing that we will be focusing on—propositional and tacit—cut across the divide of what Abel introduces as the *narrow* and *broad* conceptions of knowledge (ibid., p. 2). The narrow concept pertains to cognition that is tied to methodologically governed procedures: justification, truth, communicability, and intersubjective verifiability—the kind of knowledge paradigmatically associated with the (natural) sciences. The broad concept, by contrast, encompasses the abilities involved in adequately grasping situations, the basic competencies and skills that permeate human action, speech, thought, and perception. It follows from the assumption of plurality that we assume the distinction between propositional and tacit forms of knowing to be neither exhaustive nor sharply bounded: in an actual epistemic

practice, we are never confronted with a "single pure" form of knowledge in isolation. The phenomena, processes, and states involved in concrete episodes of knowing always already involve a fusion of different knowledge forms; it is only through subsequent heuristic reflection that we separate them into distinct components (Abel and Conant 2012, p. 33).

Second, we note that when we speak of "forms" or "types" of knowledge throughout this thesis, this terminology should not be taken to imply a set of predetermined and mutually exclusive containers into which all instances of knowing must be sorted. Following Abel, who paraphrases the notion in an explicitly Kantian register as "modes and kinds" of knowledge, we understand these categories not as rigid structures but as heuristic instruments for making the multifarious character of knowing analytically tractable (ibid., p. 6). They serve as conceptual tools that allow us to discern and articulate different dimensions of epistemic phenomena without presupposing that these dimensions correspond to ontologically separate domains.

Third, and importantly, our commitment to epistemic pluralism does not collapse into relativism. We follow Abel insofar as he emphasizes that while the plurality of knowledge forms is irreducible—we are systematically, not merely contingently, obliged to reject the view that only a single perspective suffices for addressing epistemological questions—this plurality is nevertheless subject to strict theoretical, practical, and everyday constraints (ibid., p. 10)—for, as we shall see, not everything that presents itself as knowledge actually is knowledge. The task, then, is to articulate the nature of these constraints and to specify how they guide our evaluation of what counts as genuine knowledge—a task that becomes particularly pressing when we turn to artificial cognitive systems whose epistemic capacities may not map straightforwardly onto familiar human paradigms.

Interdisciplinarity

A further methodological commitment concerns the relationship between philosophy and the empirical sciences—in our case, the field of machine learning. Abel argues that knowledge research becomes necessary precisely when the problems confronting us can no longer be defined or resolved in a purely disciplinary manner; it reveals itself, in the face of such challenges, as an inter- or transdisciplinary, reflective enterprise (ibid., p. 11). This observation applies with particular force to the question of machine knowledge. Philosophy cannot simply proceed as though the technical details of how neural networks learn, represent, and generalize were wholly irrelevant to questions about the nature

of their epistemic competence. Conversely, the field of machine learning cannot act as though there were no need to clarify the foundational concepts it employs—concepts like "knowledge," "learning," "representation," and "understanding"—all of which are inseparably tied to a longstanding history of philosophical thought.

We have already begun to address the structural problem in the current state of the discourse in this context. ML or AI research often develops independently of philosophical reflection on its key concepts, and philosophy of AI often proceeds without sufficient engagement with the technical and mathematical realities of the systems it discusses. Our thesis is therefore designed to address precisely this gap: the first analytical chapter (Chapter 2) reconstructs the technical architecture of neural networks in sufficient depth to ground the subsequent philosophical analysis, and the epistemological frameworks developed in Chapter 3 are selected and articulated with a view to their applicability to the specific technical features established in the first chapter.

Epistemic Objects and the Methodological Spectrum

The project of subjecting an artificial neural network to epistemological assessment presupposes a willingness to extend the domain of epistemological inquiry beyond its traditional boundaries. Abel makes the important point that epistemology is not a closed field of study whose themes have been settled once and for all; rather, the forms of knowledge we investigate, and the epistemic objects we consider, can and must be expanded as new challenges arise (ibid., pp. 19–20). On his account, an *epistemic object* is "[...] that toward which we direct our epistemic curiosity and attention—i.e., our knowledge-oriented attention" (ibid., p. 28). Such objects populate a broad spectrum, from elementary particles to mathematical structures to the constructions of understanding and reason. The question of what a trained neural network knows, or whether it can be said to know at all, therefore represents a new entry in this spectrum: a novel epistemic object that demands new forms of epistemological engagement.

In order to operationalize these commitments for our investigation, we thus summarize the foregoing methodological considerations into two concrete methodological guidelines: First, the recognition of epistemic plurality leads us to reject what might be summarized as *methodological reductionism* (Funk 2023, 84ff.)—the assumption that complex epistemic phenomena can be adequately captured by reduction to a single analytical framework or a single dimension of analysis. The question "Do ANNs know?" cannot be answered by appeal to technical specifications alone, nor by philosophical argument in isolation from

technical detail, nor by normative stipulation about what we choose to call "knowledge." Rather, an adequate answer requires integrating insights from machine learning theory, from classical and contemporary epistemology, and from reflection on the purposes and contexts in which we attribute epistemic predicates. Our analysis will therefore proceed by developing multiple epistemological frameworks and examining how each illuminates different aspects of ANNs' cognitive capacities.

Second, we must guard against two common fallacies that each denotes an extremum within this methodological spectrum. On the one side, there is what is commonly known as the *anthropomorphic fallacy* (Funk 2022, chap. 4.6)—the inferential error of unwarrantedly attributing human qualities (especially mental states or agency) to a non-human entity and then treating that attribution as grounds for further conclusions. In our context, this would mean that we simply infer from the fact that ANNs exhibit behaviors that, in human contexts, would manifest knowledge, that they must therefore possess knowledge in the same sense that humans do. On the other side, there is what Millière and Buckner 2024 have termed the *redescription fallacy*—the inference that something cannot possibly possess epistemic capacities simply because its operations can be "redescribed" in deflationary, mechanistic terms. In the context of ANNs, this would mean that there is no need to examine what we are about to examine simply because the material implementation of ANNs disqualifies them from epistemic assessment from the start (a priori). Our task is to navigate between these extremes and therefore to neither simply assume that behavioral similarity licenses epistemic attribution nor simply dismiss the possibility of machine knowledge on an a priori basis.

That being said, we must also emphasize that our interest and area of focus are distinctly *epistemological*. To ask what epistemic competence we are justified in ascribing to ANNs is not the same question as asking whether they possess consciousness, whether they have phenomenal experience, or whether they instantiate genuine intentionality. These are important questions in philosophy of mind, but they are not our questions. This methodological bracketing does not presume that consciousness and phenomenology are irrelevant to knowledge—that would be a substantive philosophical claim requiring argument. Rather, it reflects a strategic decision to isolate specifically epistemic dimensions of evaluation from broader questions in philosophy of mind that, however important, threaten to overdetermine our analysis.

1.3. Research Questions

The methodology we employ is therefore both *integrative* and *perspectival*. It is integrative in that it brings together technical analysis of how ANNs function with philosophical analysis of what knowledge consists in. It is perspectival in that it examines the question of machine knowledge through multiple epistemological frameworks, treating each as providing a distinct but potentially illuminating angle on the phenomenon. This approach yields a structure organized around three guiding research questions that correspond to three distinct but interconnected levels of analysis:

1. **Technical-Architectural Analysis**: How does the technical transition from symbolic, rule-based AI to subsymbolic, gradient-trained neural architectures—culminating in attention-based Transformer models for language—enable artificial neural networks to produce flexible, context-sensitive, and conversationally competent behavior? What are the key innovations in architecture, training procedures, and scaling that explain contemporary ANNs' remarkable linguistic and inferential capacities?

2. **Epistemological-Framework Analysis**: Which epistemological frameworks for propositional knowledge and for tacit knowledge are available to conceptualize and evaluate the notion of "knowing" in the context of such systems?

3. **Integrative-Evaluative Analysis**: When these epistemological frameworks are applied to ANNs as characterized through technical analysis, to what extent (if at all) do attributions of propositional knowledge and tacit knowledge succeed? Where should the relevant epistemic predicates be located—at the level of individual network components, at the level of the trained model as a whole, at the level of the model-in-interaction-with-users, or at the level of the broader sociotechnical system?

1.4. Structure of the Thesis

The investigation proceeds through three analytical stages corresponding to the research questions posed above.

Chapter 2: Artificial Neural Networks. This chapter reconstructs the transition from symbolic to subsymbolic AI. We begin with the rationalist ideal of cognition as

explicit, rule-governed symbol manipulation, tracing how Boolean algebra and predicate logic enabled systems like the General Problem Solver and the physical symbol system hypothesis. We then turn to the biological inspiration of neural networks: threshold models, Hebbian learning, and the perceptron as a first supervised learner whose limitations motivate multilayer networks. We explain how hidden layers enable learned representations, how backpropagation solves credit assignment, and why depth provides representational power. Finally, we connect deep learning to natural language processing through word embeddings, sequence models, attention mechanisms, and the Transformer architecture. The chapter concludes with CLERE (Conceptual Learner for Epistemic Reasoning Experiments), a hypothetical ANN that consolidates these technical components into explicit modeling assumptions about training, self-modification, and high-dimensional representations, serving as the concrete reference point for subsequent epistemic analysis.

Chapter 3: Epistemic Frameworks. Having established what CLERE is technically, this chapter develops the philosophical vocabulary for assessing what CLERE might know. We begin by reconstructing the justified-true-belief analysis and the historical debate about justification (rationalist a priori structure, empiricist sensory habit, Kantian synthesis). Gettier's challenge demonstrates that JTB can be true by luck, motivating the division between internalist accounts (justification fixed by the subject's internal epistemic profile—access internalism, mentalism, Bayesian credences) and externalist accounts (justification requires truth-conduciveness—process reliabilism). The second half addresses tacit knowledge: Ryle's knowing-how versus knowing-that distinction, Polanyi's theory (from-to integration, indwelling, explicability), and Collins's taxonomy (relational, somatic, and collective tacit knowledge). The chapter concludes with operationalized criteria for application to CLERE.

Chapter 4: Application to CLERE. This chapter applies the epistemic frameworks to CLERE, asking whether and under what conditions knowledge attributions can be coherently sustained. For propositional knowledge, we examine whether JTB applies to subsymbolic systems lacking doxastic standpoints, pursuing internalist reconstructions and externalist process-reliabilist approaches. For tacit knowledge, we assess CLERE through the Polanyian framework according to the from-to structure of integration, the role of indwelling, and the question of explicability: we ask whether CLERE's learned representations can plausibly be treated as subsidiary cues integrated into focal outputs, whether the relevant "participation" must be relocated to its training and deployment practices, and what interpretability can (and cannot) accomplish in rendering

tacit structure explicit. We then use Collins's distinction between relational, somatic, and collective tacit knowledge to separate contingently hidden dependencies (access, documentation, incentives) from forms of tacitness that would require embodiment or socialization into a form of life. The chapter closes by locating where epistemic predicates most plausibly attach—to the system, to its training history, or to the surrounding sociotechnical practice—while keeping in view the twin dangers of anthropomorphic projection and deflationary redescription.

2. Artificial Neural Networks

2.1. Symbolic AI

Key terms: Cartesian method, rationalism, Boolean algebra, predicate logic (quantifiers), symbolic AI, General Problem Solver, means–ends analysis, physical symbol system hypothesis.

In his attempt to derive a systematic method by which to avoid contradictions in the sciences and generate well-founded knowledge, Descartes presents four precepts that can be summarized as follows: (1) Accept as true only that which is undoubtedly certain; (2) divide each problem into as many smaller parts as necessary for its adequate solution; (3) assume that there is a solution to all complex problems by "[...] assigning in thought a certain order even to those objects which in their own nature do not stand in a relation of antecedence and sequence [...]"; and (4) make enumerations sufficiently complete and reviews sufficiently general to ensure that nothing was omitted (Descartes 1985, p. 120). From an epistemic point of view, these rules articulate an ideal of cognition as explicit, stepwise, and rule-governed: knowledge should rest on what can be grasped with clarity and distinctness (rule 1), complex questions should be rendered tractable by analysis into simpler components (rule 2), reasoning should be reconstructible as a determinate sequence of inferential moves (rule 3), and the resulting chain should be checkable for gaps (rule 4). Descartes illustrates this strategy with reference to Euclidean geometry, thereby presenting mathematics as a model for what it would mean to have secure cognition: a method that, in principle, can be followed, inspected, and corrected.[1]

This Cartesian notion stands paradigmatic for seventeenth-century rationalism and its tendency to frame cognition as a system of rules and operations rather than as a merely associative or experiential flow. The rationalist ambition—to explain how knowledge can be necessary, systematic, and resistant to error—will be taken up explicitly in the next chapter (Section 3.1), where the focus shifts toward the epistemological role such ideals of method play in accounts of justification, inference, and knowledge. For the present purposes, however, what matters is how this ideal of formalizable reasoning prepares the ground for the emergence of *symbolic AI*:[2] the

[1] See Coreth and Schöndorf 1983.

[2] Another important philosophical precursor in this regard is Leibniz, who radicalized the rationalist program by envisioning not merely a method of reasoning but a universal symbolic language—the *characteristica universalis*—in which all concepts could be represented by formal signs, together with a *calculus ratiocinator*, a mechanical calculus by which truths could be derived through rule-governed manipulation of those signs alone (Leibniz 1969, pp. 221–228). Where Descartes proposed procedural rules for the thinking subject, Leibniz thus anticipated the prospect of externalizing reasoning itself into a symbolic calculus that could, in principle, be carried out by a machine—an idea that constitutes one of the most direct philosophical precursors of symbolic AI.

idea of interpreting thinking as symbol processing and transferring it to machines with the help of increasingly complex logical-mathematical programming (Gramelsberger and Sigwart 2025, p. 63).

"Good old-fashioned artificial intelligence," as John Haugeland famously named it (Haugeland 1993), started to take shape in the nineteenth century as logicians and mathematicians like George Boole and Gottlob Frege drove the necessary progress in formal logic. With *The Laws of Thought* (1854), Boole extended the Aristotelian tradition by showing how logical relations can be treated algebraically: instead of classifying valid moods and figures, one can calculate consequences by transforming symbolic expressions according to general rules.[3] Frege's introduction of quantification and his function-argument analysis in the *Begriffsschrift* (1879) then yields a markedly more expressive formalism—predicate logic—capable of representing generality, relations, and inference patterns in a way that later becomes foundational for logic-based approaches to cognition and computation.[4]

These progressions in formal logic provided the necessary tools to elevate the theoretical concepts developed by seventeenth-century rationalists such as Descartes and Leibniz to the first practical experiments of "calculating thought"—as Leibniz put it (Kulstad and Carlin 2020)—in the twentieth century. By mapping Boolean algebra to relay and switching circuits—i.e., discovering how symbol manipulation can be realized physically[5]—problems represented by formal logic could be automated at scale. Using this new and powerful way to calculate in much larger dimensions, Newell, Simon, and Shaw invented the "General Problem Solver" (GPS) (Newell and Herbert A. Simon 1988): the program which, together with its predecessor "The Logic Theorist" (LT) (Newell and Herbert A. Simon 1956), is usually referred to as "the birth of artificial intelligence" (S. Russell and Norvig 2016, p. 19). The main characteristic of the GPS, the property that warranted the designated term "intelligence," was that it could—being the

[3] A standard illustration is the Aristotelian syllogism AAA–1 ("Barbara") (Lagerlund 2022): *All A are B; all B are C; therefore all A are C*. Under Boole's encoding of class inclusion via equations (e.g. $a = ab$ for "all A are B", with multiplication representing intersection), the premises $a = ab$ and $b = bc$ yield $a = ac$ by substitution and associativity.

[4] The key difference from Boole-style propositional treatment is that Frege's notation can make the *internal structure* of statements explicit (Cook 2024). Instead of treating sentences as unanalyzed atoms (e.g. P, Q), one can represent predication and relations by applying a predicate to an argument (or several arguments), e.g. $Human(x)$, $Mortal(x)$, or $Loves(x, y)$. Quantifiers then express generality and existence directly: $\forall x\,(Human(x) \rightarrow Mortal(x))$ formalizes "All humans are mortal," and $\exists x\, Loves(x, y)$ can express that someone loves a given person y (ibid.). This additional expressive power is precisely what later makes first-order logic suitable for knowledge representation and rule-based inference in symbolic AI (e.g. representing facts and general rules that can be instantiated to derive conclusions about particular individuals).

[5] See Shannon (1938, pp. 713–723), an abstract of Shannon's MIT master's thesis, for the classic demonstration that Boolean algebra can be used as a calculus for analyzing and simplifying relay and switching circuits—i.e., how logical symbol manipulation can be implemented in physical switching hardware.

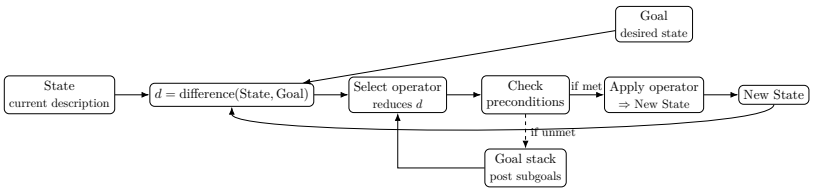


Figure 2.1.: GPS means-ends analysis.

Source: Designed by the author and implemented using generative AI tools; see Appendix A.1.

first artificial entity to do so—separate its domain knowledge (the knowledge of the rules) from its solving strategy, which enabled one solver to work across different domains (Newell and Herbert A. Simon 1988, 118ff.).

The way it worked was that it had a *state* (the symbolic description of "where it is"), a *goal* (the symbolic description of "where it wants to be"), and operators (domain actions that can transform one state into another) (ibid., 114ff.). It then moved in a given task environment by selecting an operator that best *reduces the difference* between the current state and the goal—a problem-solving methodology known as "means–ends analysis" (Herbert Alexander Simon 2008, p. 121). If the operator's preconditions were not satisfied, the solver posted them as subgoals on a "goal stack" and recursively worked to achieve them (see Figure 2.1).

Unlike its predecessor LT, GPS was designed from the start to imitate human problem-solving protocols. Newell and Simon would set up experiments where they instructed participants to solve a given problem while thinking aloud, recorded that process, and then ran the GPS on that transcript to produce a comparison (Newell and Herbert A. Simon 1988, 111ff.). The success of the new methodology which enabled GPS to cross domains eventually led Newell and Simon to formulate the famous *physical symbol system hypothesis* (Newell and Herbert A. Simon 1976, p. 116), which states that "a physical symbol system has the necessary and sufficient means for general intelligent action" (ibid., p. 116). Thus, in theory, so they claimed, any problem that can be formulated as a set of well-formed formulas and that allows for the state-goal-operator structure can be solved by a symbolic AI program like the GPS.

2.2. Subsymbolic AI

Key terms: Neurons, all-or-nothing principle, Boolean networks, Hebb's learning rule, machine learning.

As symbolic AI operates under the assumption that human cognition is best described and modeled as rule-based symbol manipulation—a claim put to the test in subsequent chapters (Chapters 3 and 4)—the paradigm that eventually led to AI systems such as large language models (LLMs) is based on a different idea: the biological nervous system. What we know about biological nervous systems is that—on some level of abstraction—they consist of basic computational units called nervous cells or *neurons* (see Figure 2.2) (Hofmann 2022, p. 8). Neurons are connected by nerve fibers, which allows for communication between neurons via relatively fast ($\approx$ 100m/s) and long-range (up to $\approx$ 1m) electrical impulse signal transmission.[6] The specific trait of neurons which led to the first big breakthrough in modeling their functions is that their communication is based on an *all-or-nothing principle* (Hofmann 2022, p. 10).

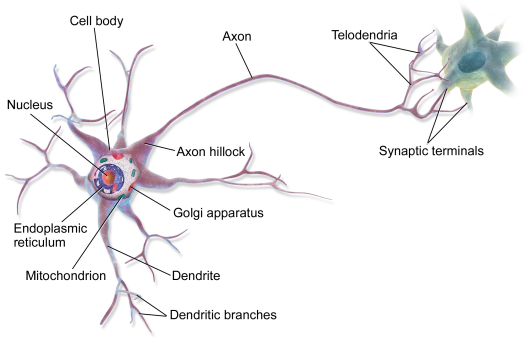


Figure 2.2.: Anatomy of a neuron.
Source: BruceBlaus 2013.

A neuron's action potential—its *spike*—is a binary event: only when synaptic inputs depolarize the membrane at the axon initial segment (the spike "trigger zone") past its threshold does a full spike (action) occur and propagate down the axon (the main transmission pathway for signals in the nervous system); if the threshold is not reached, no spike is generated (ibid., p. 10). In their paper *A Logical Calculus of Ideas Immanent in Nervous Activity* McCulloch and Pitts (1943) thus deduce: "Because of the 'all-or-none' character of nervous activity, neural events and the relations among them can be treated by means of propositional logic [...]" (ibid., p. 115). Specifically, as *Boolean circuits*, i.e., logical gates that take 0/1 inputs and compute 0/1 outputs.

[6] Neurons also communicate via slower, diffuse chemical signaling and via graded changes in membrane potential in addition to all-or-nothing spikes. For a standard textbook overview of these mechanisms, see Kandel et al. (2021).

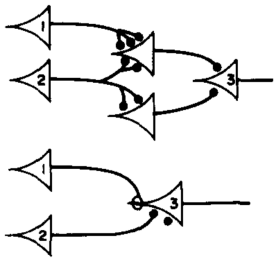


Figure 2.3.: McCulloch-Pitts Boolean-circuit scheme.

Source: McCulloch and Pitts 1943, p. 130.

In the original scheme (Figure 2.3), each triangle (1, 2, 3) represents a neuron—the rightmost being the output of that circuit. The lines going from each neuron are either excitatory synapses, which increase the chance of a spike (black dots), or inhibitory synapses, which decrease the chance of a spike (plain ones). In a model with relative inhibition,[7] a neuron has n inputs, a signature $\sigma \in \{-1, 1\}^n$ (excitatory vs. inhibitory synapses), and a threshold $\theta \in Z$. It then computes its action (output) as a Boolean function (Hofmann 2022, p. 18)

$$f(x;\, \sigma, \theta) \;=\; \begin{cases} 1, & \text{if } \sum_{i=1}^{n} \sigma_i \, x_i \;\geq\; \theta, \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

As we can see, McCulloch & Pitts's idea of modeling nervous systems as logical gates still shares an important feature with symbolic AI: each neuron corresponds to a well-formed logical condition on its predecessors, and the system as a whole does not yet allow for memory or learning—it is still static in this sense. The pivotal theory that changed this was introduced six years later (1949) by the psychologist Donald Hebb.

Hebb observed that if a presynaptic cell and a postsynaptic cell are repeatedly co-active, the synapse—the connection between them—strengthens (D. O. Hebb 1949). Thus the famous postulate: *Neurons that fire together, wire together* (Shatz 1992, p. 64). With this Hebbian learning rule, as it is generally known, the fixed synaptic signs of the McCulloch & Pitts model are replaced by real-valued synaptic *weights* $w_{ij} \in R$ that can change as the model progresses. In its simplest form, a Hebbian learning rule for binary neurons would thus be

$$\Delta w_{ij}^t \propto x_i^t x_j^t, \tag{2.2}$$

[7] There are generally two ways of modeling inhibition (McCulloch and Pitts 1943, p. 123): absolute (veto) inhibition, which means that if any inhibitory input is 1, the neuron must be 0 at the next tick—no matter how many excitatory inputs are on—and relative inhibition (the modern approach), where inhibitory inputs only reduce the chance of firing by contributing -1 to the sum.

i.e., the connection $j \to i$ is strengthened at time t precisely when the presynaptic unit j and the postsynaptic unit i are simultaneously active (Hofmann 2022, p. 28). Although, as a mathematical model, this (naïve) form of the Hebbian rule is not yet of practical use—it simply strengthens any pair of units that happen to be co-active[8]—the important conceptual breakthrough toward *machine learning* is the idea that a computational system can *modify itself* (Shalev-Shwartz and Ben-David 2014, pp. 21–22): the connection strengths w_{ij} are not fixed design choices but parameters that change in response to experience (new data). In other words, with Hebb, learning becomes a rule that maps activity statistics (and, when available, feedback about success or failure) into parameter updates. This shift—from static Boolean circuits to adaptive, parameterized networks—opened the door to practical learning algorithms.

2.3. Perceptron

Key terms: Decision boundary, activation function, learning rate, perceptron convergence theorem, linear separation, XOR.

One of the first architectures that combined the idea of using a simple threshold function (Equation (2.1)) and a learning rule (Equation (2.2)) into a fully operational neural network was Frank Rosenblatt's *perceptron* (F. Rosenblatt 1958). To illustrate the idea of the perceptron, consider Figure 2.4. It shows a two-dimensional feature space—that is, a Cartesian plane in which each axis represents one measurable attribute (feature) of an example. A *binary classification* task asks us to assign each example to one of two classes; we denote the class label by $y \in \{-1, +1\}$. A labeled example is therefore a point x in the plane together with its label y. The finite set of labeled examples used to learn is the training set, written $D = \{(x_i, y_i)\}_{i=1}^{N} \subset R^2 \times \{-1, +1\}$.

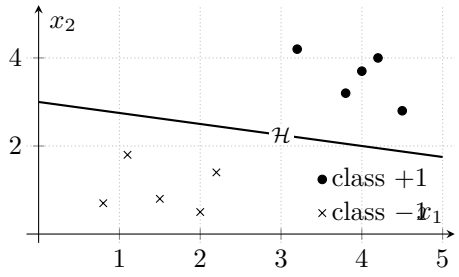


Figure 2.4.: Binary classification.

Source: Designed by the author and implemented using generative AI tools; see Appendix A.1.

[8] Without additional constraints, the naïve Hebbian model (1) drifts to saturation, i.e., the weights grow unbounded, (2) is blind to the task (since there is no notion of when a synaptic unit fires for the "right" reasons), and (3) confounds mere baseline activity with meaningful co-variation (the rule strengthens synapses purely because both are often "on," not because they co-vary beyond chance) (Hofmann 2022, p. 28).

The goal of the perceptron is to construct a *decision boundary* (Hofmann 2022, p. 21) that separates the feature space into the two classes based on the training data D. We denote this boundary by $\mathcal{H}$:

$$\mathcal{H} = \{\, x \in R^2 \mid w_1 x_1 + w_2 x_2 = \theta \,\}. \tag{2.3}$$

Combining the concept of adjustable weights $w = (w_1, w_2)$—which quantify how strongly each coordinate of $x = (x_1, x_2)$ contributes—with the McCulloch & Pitts threshold function (Equation (2.1)) used here as the activation (see Figure 2.5), the perceptron computes for each example x_i the weighted sum

$$z_i = w_1 x_{i1} + w_2 x_{i2},$$

and predicts

$$\hat{y}_i = \begin{cases} +1, & \text{if } z_i \geq \theta, \\ -1, & \text{otherwise.} \end{cases}$$

Thus, points with $w_1 x_1 + w_2 x_2 > \theta$ fall on one side (predicted $+1$), and those with $w_1 x_1 + w_2 x_2 < \theta$ on the other (predicted -1) (ibid., 21ff.).

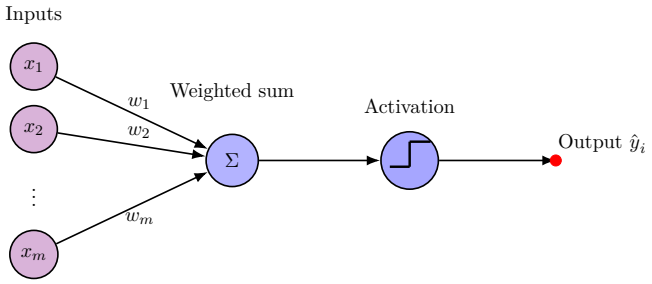


Figure 2.5.: Perceptron architecture.
Source: Adapted from F. Rosenblatt 1958.

To make these adjustments gradual and controllable, we introduce a *learning rate* $\eta > 0$ (ibid., p. 97). The learning rate is a small positive number that scales each weight change: a large η moves the boundary $\mathcal{H}$ quickly but can overshoot; a small η moves it slowly but more steadily. As an extension of Hebb's idea (Equation (2.2)), the perceptron updates the weights only when an example is misclassified. Concretely, for labels $y_i \in \{-1, +1\}$ and score $z_i = w_1 x_{i1} + w_2 x_{i2}$, if $y_i(z_i - \theta) \leq 0$ the update

$$w \leftarrow w + \eta\, y_i\, x_i \quad \text{(and, optionally,} \quad \theta \leftarrow \theta - \eta\, y_i)$$

nudges $\mathcal{H} = \{x : w_1 x_1 + w_2 x_2 = \theta\}$ toward correcting that mistake.[9] In this way, learning

[9] Equivalently, one can absorb the threshold into a *bias* parameter by setting $b := -\theta$ and augmenting the input with a constant feature $x_0 = 1$; then the boundary is $\{x : w^\top x + b = 0\}$ and the optional threshold update above corresponds to $b \leftarrow b + \eta\, y_i$ (Hofmann 2022, p. 45).

iteratively moves the decision boundary until it separates the two classes (given that they are linearly separable) in the given data or some other stopping criterion is met.

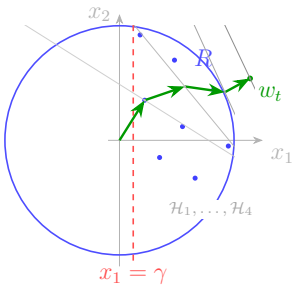


Figure 2.6.: Geometric visualization of the perceptron convergence theorem.
Source: Designed by the author and implemented using generative AI tools; see Appendix A.1.

As we can see, although the perceptron, at its core, consists only of the combination of two simple functions, it is capable of performing real-world classification tasks, i.e., learning patterns. Let us assume a dataset D where (1) all inputs are bounded by a radius R so that $\max_{(x,y)\in D} \|x\|_2 = R$ (for simplicity, Figure 2.6 shows the negative examples as reflected through the origin), (2) a margin so that every point sits at least γ away on the correct side along its normal, and (3) a unit vector w^* that is capable of linearly separating the data. In this case, assuming unit-norm data ($R = 1$), the *perceptron convergence theorem* (Frank Rosenblatt 1962, pp. 99–101) as proposed by Rosenblatt proves that "[...] the perceptron converges in at most $\lfloor \gamma^{-2} \rfloor$ update steps on any γ-separable sample" (Hofmann 2022, p. 23), for any learning rate and any method of sampling from the dataset.[10]

That being said, if the data are not linearly separable—which is the case for most real-world problem settings—the perceptron in its current form cannot find a solution. The simplest and most well-known example illustrating this limitation is the XOR ("exclusive OR") function (Figure 2.7) (Hofmann 2022, p. 24). As can be clearly seen, there is no linear decision boundary (Equation (2.3)) capable of separating the two classes.[11]

[10] See Hofmann 2022, 23ff. for an outline of the proof.

[11] One might note here that the XOR pattern can be made linearly separable after a suitable feature transformation (e.g., by mapping the inputs into a higher-dimensional feature space (Shalev-Shwartz and Ben-David 2014, p. 217)). However, such transformations are not always available and, even when they are, they are typically not straightforward to identify. This is precisely why one needs models that can *learn* more flexible, non-linear decision boundaries, such as multi-layer perceptrons.

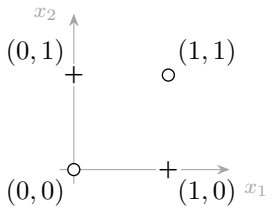


Figure 2.7.: XOR geometrically.

Source: Designed by the author and implemented using generative AI tools; see Appendix A.1.

2.4. Multilayered Networks

Key terms: forward pass, backpropagation, gradient, loss function, cross-entropy, logits, softmax, activation function, credit assignment, hidden layers, representation learning, non-linear decision boundary, optimization (gradient descent), universal approximation theorem.

The key to overcoming the limitations of the single-layer perceptron (SLP)—as the name already suggests—is to add hidden layers (see Figure 2.8). Continuing the idea of modeling a network inspired by biological nervous systems, multilayer neural networks—e.g., multilayer perceptrons (MLPs)—extend the SLP (Figure 2.5) by feeding the output of one layer into the next. The purpose of these layered transformations is to build representations that emphasize or de-emphasize different aspects of the input features so that the classes become (approximately) linearly separable (Hofmann 2022, 44ff.).

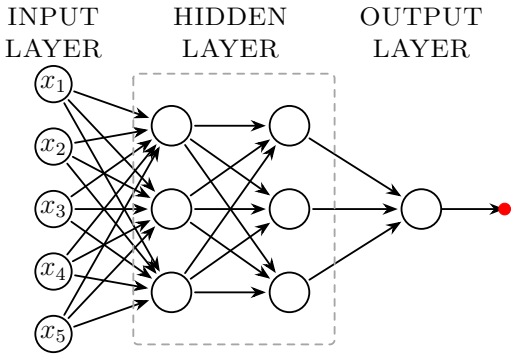


Figure 2.8.: Schematic multilayer perceptron with two hidden layers.

Source: Designed by the author and implemented using generative AI tools; see Appendix A.1.

One of the first tasks that arises with this expanded SLP architecture is to find an efficient way to determine how each part of the multilayer transformation contributes to the final prediction—i.e., how to optimize the parameters toward the correct solution. In the SLP

this is straightforward: with only one layer, whenever the algorithm makes a mistake, the corresponding weights are updated. However, once we introduce more than one layer, two issues appear (Shalev-Shwartz and Ben-David 2014, 184ff.): (1) the step decision is non-differentiable, so we cannot apportion credit to internal weights, and (2) correctly classified points provide no training signal. MLPs address these issues in two ways: (1) by quantifying the discrepancy between predictions and targets with a scalar, differentiable *loss function*, and (2) by propagating that loss backward through the network to update the parameters (*backpropagation* (ibid., p. 277)).

What we want to quantify with the loss function is how confidently the model makes its decision and, accordingly, reward high confidence when correct, gently correct near-misses, and strongly penalize confident mistakes. Take *cross-entropy loss* (Hofmann 2022, p. 58) as an example: since Shannon's characterization of self-information (Shannon 1948) implies that if an event occurs with probability p, its surprisal is $S(p) = -\log p$,[12] we get the binary cross-entropy (BCE) loss for a model that outputs a probability $p \in (0, 1)$ for class "1" where the true label is $y \in \{0, 1\}$ as

$$\mathcal{L}_{\mathrm{BCE}}(p, y) \;=\; -\Big(y \,\log p \;+\; (1 - y)\,\log(1 - p)\Big). \tag{2.4}$$

And equivalently, for multiclass problems, where the MLP outputs a probability vector $p = (p_1, \ldots, p_K)$ with $\sum_{k=1}^{K} p_k = 1$, the cross-entropy per example is

$$\mathcal{L}_{\mathrm{CE}}(p, y) \;=\; -\sum_{k=1}^{K} y_k \log p_k \;=\; -\log p_c. \tag{2.5}$$

where y is the true label encoded as a one-hot vector (i.e., $y_c = 1$ and $y_k = 0$ for $k \neq c$), and p_c denotes the model's predicted probability for the correct class c (Hofmann 2022, p. 58). In this way, we now have a scalar that measures how far the predictions are from the target.

The next step in assigning credit is to propagate this scalar back through the network to compute how an infinitesimal change to the parameters would affect that loss (ibid., 75ff.). To illustrate

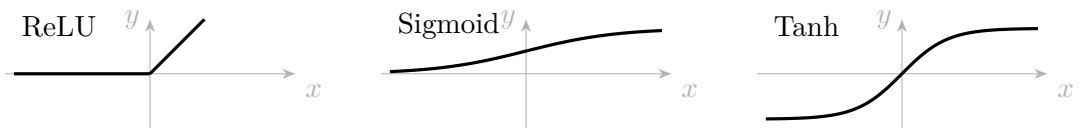
[12] According to Tribus and McElroy 1971, a "surprise score" $S(p)$ for an event of probability p should (i) decrease when p increases (if $p_1 < p_2$ then $S(p_1) > S(p_2)$), (ii) be zero when $p = 1$ (if something is certain, there is no surprise), and (iii) add for independent events A, B,

$$S\big(P(A \cap B)\big) = S\big(P(A)P(B)\big) = S\big(P(A)\big) + S\big(P(B)\big).$$

Since seeing two independent unlikely events in a row should be more surprising, and the total surprise should be the sum of each one's surprise. A standard and particularly convenient way to implement this "multiplication-to-addition" requirement is to use a logarithmic form, thus leading to:

$$S(p) = -\,k\,\log p, \qquad k > 0.$$

the idea, let us start from where we left off with the SLP: the forward pass. Just as with the SLP, the input is fed through hidden layers and undergoes the same affine transformations $z^\ell = W^\ell a^{\ell-1} + b^\ell$; but instead of applying the binary threshold function (Equation (2.1)), an MLP uses a differentiable activation such as ReLU ($\max(0, x)$), sigmoid ($1/(1+e^{-x})$), or tanh ($(e^x - e^{-x})/(e^x + e^{-x})$) (Shalev-Shwartz and Ben-David 2014, p. 269).



After each hidden layer, we obtain the activation $a^\ell = \varphi(z^\ell)$. After the final hidden layer, a last affine map produces the *logits* $z^L = W^L a^{L-1} + b^L$, which we convert into a probability score (sigmoid for binary, softmax for multiclass):

$$\text{Binary (sigmoid)} \qquad p = \sigma(z^L) = \frac{1}{1+e^{-z^L}} \in (0,1), \tag{2.6}$$

$$\text{Multiclass (softmax)} \qquad p_k = \frac{e^{z_k^L}}{\sum_{j=1}^K e^{z_j^L}}, \qquad \sum_{k=1}^K p_k = 1. \tag{2.7}$$

That forms the cross-entropy loss as shown above (Equation (2.4) or (2.5)), yielding a single scalar per example to minimize (see Figure 2.9).

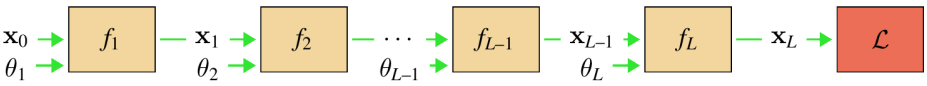


Figure 2.9.: Computational graph view of forward pass.
Source: Torralba, Isola, and Freeman 2026.

To learn from this loss—i.e., to figure out how to change the weights so the loss decreases—we propagate the output error obtained by comparing the model's probabilities to the target back through the network. Mathematically speaking, we set

$$\delta^L = \frac{\partial \mathcal{L}}{\partial z^L} = \begin{cases} p - y, & \text{(sigmoid + binary cross-entropy)} \\ p_k - y_k \ \text{ for } k = 1, \ldots, K, & \text{(softmax + multiclass cross-entropy)} \end{cases}$$

and then apply the chain rule layer by layer. For each hidden layer $\ell = L-1, \ldots, 1$,

$$\delta^\ell = (W^{\ell+1})^\top \delta^{\ell+1} \odot \varphi'(z^\ell),$$

where $\odot$ denotes element-wise multiplication and φ' is the derivative of the activation. From these error signals we read off the parameter gradients:

$$\frac{\partial \mathcal{L}}{\partial W^{\ell}} = \delta^{\ell} \, (a^{\ell-1})^{\top}, \qquad \frac{\partial \mathcal{L}}{\partial b^{\ell}} = \delta^{\ell},$$

continuing this recursion until the first layer is reached (see Figure 2.10).

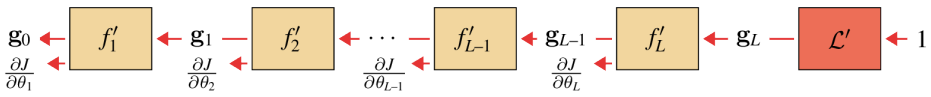


Figure 2.10.: Computational graph view of backpropagation.
Source: Torralba, Isola, and Freeman 2026.

In short, by introducing a differentiable loss and backpropagation, we turned the perceptron's mistake-driven rule into a continuous optimization procedure that assigns credit to every weight and bias term and updates them via gradients (Torralba, Isola, and Freeman 2026). This lets an MLP learn internal representations that progressively reshape the input so that originally non-separable classes become separable at some hidden layer. To illustrate that, consider the XOR problem (Figure 2.7) from above: using a multilayered network with just two hidden layers, we can now separate the non-linear XOR by composing two linear decision boundaries (Figure 2.11).

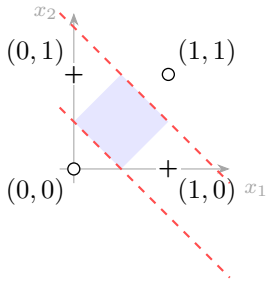


Figure 2.11.: XOR with MLP decision boundary.
Source: Designed by the author and implemented using generative AI tools; see Appendix A.1.

This ability of multilayered networks to overcome the SLP's constraints by stacking affine layers with pointwise non-linear activations and training all parameters via a differentiable loss and backpropagation leads to the *universal approximation theorem* (UAT) (Hornik, Stinchcombe, and White 1989) that states: an MLP with a single hidden layer and a fixed non-linear, non-polynomial activation can approximate any "well-behaved" input-output rule—formally, any continuous function on a bounded (compact) input set—arbitrarily well. More precisely,

let $K \subset R^n$ be compact and $f \in C(K)$. Then, for every $\varepsilon > 0$ there exist a width m and parameters $\{\alpha_j, w_j, b_j\}_{j=1}^m$ such that the one-hidden-layer network

$$x \longmapsto \sum_{j=1}^{m} \alpha_j \, \phi(w_j^\top x + b_j)$$

satisfies

$$\sup_{x \in K} \Big| \, f(x) - \sum_{j=1}^{m} \alpha_j \, \phi(w_j^\top x + b_j) \, \Big| \; < \; \varepsilon,$$

where ϕ must be sufficiently expressive; in particular, universality holds for standard choices such as sigmoids and tanh, and, more generally, for any *non-polynomial* activation including ReLU (Hofmann 2022, p. 62). However, it is important to note that UAT is a *representational* guarantee only: it does not specify how large the network must be to reach a prescribed error, whether gradient-based optimization will find the requisite parameters from finite, noisy data, how long training will take, or how well the learned model will generalize to unseen inputs. It only tells us that the parameters exist in theory, but, as we shall see, the success of actually finding them in practice depends on data engineering factors such as regularization, optimization, data quality, and so on.

2.5. Deep Learning and Natural Language Processing

Key terms: scale (data/compute), NLP, discrete sequences, long-range dependencies, context dependence, Zipf's law, one-hot encoding, sparsity, distributional hypothesis, word embeddings, Word2Vec, GloVe, cosine similarity.

At its core, the term "deep learning" (DL) is the label most commonly used to describe the learning process associated with a multilayered neural network structure (Figure 2.8) such as we have just introduced.[13] In that regard, we have already covered the essential concepts relevant to our level of understanding: representation by layers, loss-based training, and gradient-based credit assignment. What has changed since the late 20th century, i.e., led to the major breakthroughs of the last two decades, is not so much these methodological foundations—the math behind the algorithms—but the scale and stability by which they can be applied. Most notably, the recent success that led to DL models such as ChatGPT can be attributed to (1) the

[13] The term *deep learning* is used somewhat loosely, but in a widely cited, standard sense it refers to learning in models composed of multiple processing layers that learn representations with multiple levels of abstraction (LeCun, Bengio, and Hinton 2015). A complementary textbook characterization understands deep learning as learning hierarchical representations (many "levels" of concepts) from data (Goodfellow, Courville, and Bengio 2016). In this chapter, the term is used in this broad sense: multilayer neural networks trained via gradient-based optimization.

availability of data,[14] (2) the increase in computational power (parallel computing, commodity GPUs, modern software tools), and (3) the training/architecture refinements (non-saturating activations, batch normalization, residual connections) (Goodfellow, Courville, and Bengio 2016, pp. 19–21); see also (LeCun, Bengio, and Hinton 2015, p. 436). The last step of the technical evolution that is thus left for us to cover is the application of ANNs to the field of natural language[15]—since, as we shall see in the next chapter, this is an important domain where the epistemic status of machines is debated (see Chapter 3).

From a technical point of view, language holds a specific set of obstacles that, until very recently, were thought to be one of the main reasons why artificial learners would never master language as humans do (Bar-Hillel 1960; National Research Council (U.S.). Automatic Language Processing Advisory Committee 1966). In particular, text—the primary substrate processed by language models (LMs)—is (1) discrete and order-sensitive over variable-length sequences ("philosophy" $\neq$ "yphosolihp"); (2) exhibits long-range dependencies, so relevant cues can be separated by many tokens ("The article I read yesterday on the subject of post-quantum cryptography that Professor X ... was excellent"); (3) is context-dependent and compositional, so a single token can shift meaning with grammar and discourse ("bank" as riverbank vs. financial institution); (4) follows heavy-tailed frequency distributions—some words like "the," "and," "is" appear much more than others (Zipf's law (Zipf 1949))—inducing sparsity and out-of-vocabulary effects—i.e., some words never appear in the training data and are therefore unknown to the model; and (5) underwrites many sequence-to-sequence tasks (translation, summarization, dialogue) in which models need to first form an internal understanding of the input and then generate a new sentence that keeps the same meaning, even if the wording and order are different (Jurafsky and Martin 2026; Manning and Schütze 1999; Bengio, Simard, and Frasconi 1994).

It becomes apparent that it would be difficult to apply the simple unmodified MLP directly to this kind of data: let us say we would use the same one-hot encoding strategy introduced in the previous section (Section 2.4) where each word is represented as a vector with a 1 at the index assigned to that specific word (1: "cat" $\rightarrow$ [1, 0, 0 ...]). For a small vocabulary with only 10k words, this would mean every vector has length 10k and contains only one relevant element—i.e., is huge and extremely sparse. Furthermore, there is no notion of similarity between words; "cat" is just as far from "kitten" as it is from "thermodynamics." The numbers do not reflect any meaning, and the model cannot learn something about "kitten" from "cat"; it might as well be "cat" $\rightarrow$ "thermodynamics." In addition, the MLP as we have introduced it so far expects a fixed-size input ($Wx + b$) and thus cannot handle variable-length input—nor

[14] See Section 2.10 for a more precise account of the term.

[15] The term "natural language" is used to distinguish the organic use of words and grammar in human communication from formal languages such as first-order logic, mathematical notation, and programming languages. See Jackson and Moulinier 2007.

can it infer order between words in a sentence ("dog bites man" ⇔ "man bites dog") or capture long-term dependencies.

In order to handle textual data, we therefore need a way to represent words in a dense data structure that is capable of holding information on semantic relations. The first breakthrough in this context came with the *distributional hypothesis* (Z. S. Harris 1954): the assessment that semantically related words tend to appear in similar contexts—that is, as Firth 1957 famously put it, "a word is characterized by the company it keeps." This insight provided the linguistic foundation that allowed for a geometric interpretation of semantic relations.

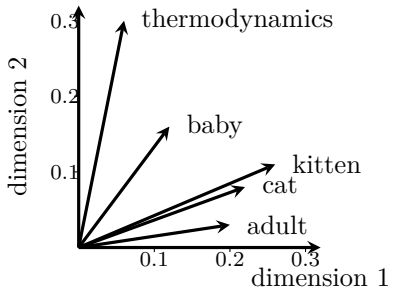


Figure 2.12.: Semantic vector space.

Source: Designed by the author and implemented using generative AI tools; see Appendix A.1.

Rather than representing each word as a sparse one-hot vector, models like *Word2Vec* (Mikolov et al. 2013) or *GloVe* (Pennington, Socher, and Manning 2014) therefore learned to map words into dense, low-dimensional vector spaces (Figure 2.12) that could be operationalized using linear algebra. These embeddings were learned by predicting words from their surrounding context (or vice versa), thereby capturing statistical regularities of word co-occurrence. In such spaces, "cat" and "kitten" would thus have vectors pointing in similar directions, while "thermodynamics" would be far away, allowing for algebraic calculations such as proximity via dot product or cosine similarity[16] that make it possible to derive analogies such as $\mathbf{v}_{\text{cat}} - \mathbf{v}_{\text{adult}} + \mathbf{v}_{\text{baby}} \approx \mathbf{v}_{\text{kitten}}$. This approach transformed the sparse V-dimensional one-hot vectors (where V is vocabulary size) into dense d-dimensional embeddings (where $d \ll V$), making the representation both more efficient and more meaningful[17] (Peters et al. 2018).

[16] $\cos(\mathbf{v}_a, \mathbf{v}_b) = \frac{\mathbf{v}_a \cdot \mathbf{v}_b}{\|\mathbf{v}_a\| \, \|\mathbf{v}_b\|}$.

[17] However, as we shall see, this kind of word embedding still suffers from a fundamental limitation: each word has exactly one vector representation regardless of meaning. "Bank" would still have the same embedding whether it appeared in "river bank" or "investment bank."

2.6. RNN, LSTM and Encoder–Decoder NLP

Key terms: variable-length sequences, word order sensitivity, recurrent state as memory, vanishing/exploding gradients, gating (selective retention/update), long-term vs. short-term memory, sequence-to-sequence translation, fixed-size context bottleneck.

The next challenge in order to apply an ANN to natural language was to overcome the limitations of the fixed input size while simultaneously figuring out a way to preserve information about word order. The first step in this direction was achieved with the introduction of Recurrent Neural Networks (RNNs). The main idea of an RNN is to introduce a hidden state parameter h_t that acts as the *memory* of the network (see Figure 2.13) (Hofmann 2022, p. 130). As the input x_t gets passed through the RNN, the weight parameters are updated with regard to the hidden state that holds the information from the previous time step. This makes it possible for the network to handle sequential input values where order matters. In contrast to an MLP, the input is therefore not a single fixed vector but an *observation sequence* $x_{1:T} = (x_1, \ldots, x_T)$, where each $x_t \in R^d$ is the (embedded) representation of the token at position t (Goodfellow, Courville, and Bengio 2016, p. 368).

To illustrate this idea, let us unroll an RNN across three steps: yesterday $(t-1)$, today (t) and the prediction for tomorrow $(t+1)$. At each step we use the same small network cell as before: the input x_t is multiplied by a learned weight (shown as $\times W_1$), we add a learned offset b_1 (the *bias*), and we pass the result through a simple nonlinearity (the *activation*). The output of this activation, labeled y_t, is the cell's *hidden state* h_t—the piece of memory that summarizes everything the cell has seen up to and including time t (ibid., 372ff.).

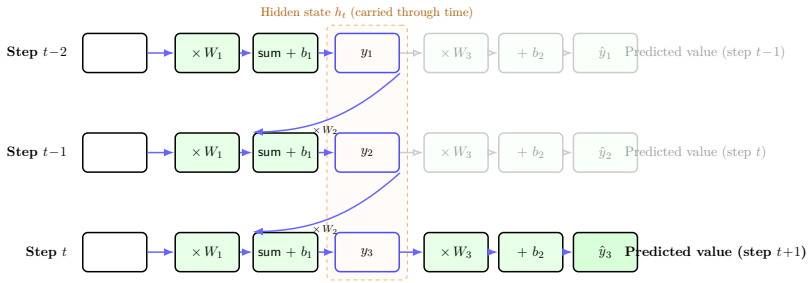


Figure 2.13.: RNN unrolled in time.
Source: Adapted from Starmer 2025.

Instead of using the post-activation value y_t to generate an output directly—as we would in a non-recurrent network—the RNN sends this state forward to the next time step,

where it is combined with the new input at that time x_t. Since *the same parameters* $(W_1, W_2, W_3, b_1, b_2)$ *are reused for all three steps*, an input such as ("dog", "bites", "man") produces a different hidden state sequence than ("man", "bites", "dog") and therefore a network that is sensitive to the input order over time (Goodfellow, Courville, and Bengio 2016, 372ff.).

Although the idea of introducing a feedback loop solves the sequence and order issue, this setup—if we trained it exactly as is—runs into a problem: repeating the same transformation many times is inherently unstable (Pascanu, Mikolov, and Bengio 2013). When an RNN is unrolled for just 50 steps (which is very little in practice), one is effectively applying the hidden state about 50 times in a row. This means that even a tiny per-step amplification gets huge: if the per-step effect is a bit smaller than 1, the overall influence shrinks exponentially (e.g., $0.5^{50} \approx 9 \times 10^{-16}$); if it is a bit larger than 1, it blows up (e.g., $1.5^{50} \approx 6 \times 10^{8}$). In training, this phenomenon is known as the *vanishing/exploding gradient* problem (ibid., p. 2): As we know, the gradient of backpropagation at an early step is a product of the recurrent weight map and the activation derivatives from every later step (see Section 2.4). Because the derivatives of tanh/sigmoid are ≤ 1 (and often much smaller), chaining them over dozens of steps typically drives the gradient toward zero—or, where the effective per-step factor exceeds 1, yields exploding gradients. Therefore, the learning signal from a word far back in the sequence would barely reach the parameters that processed it, so the model struggles to learn long-range relations (e.g., linking "article … excellent" across a long clause).

The solution to this problem is to adapt the RNN to hold *long- and short-term memory* (LSTM) (Hochreiter and Schmidhuber 1997). As we can see in Figure 2.14, an LSTM has two paths: a *long-term memory* (LTM, green line) that runs across the top and a *short-term memory* (STM, pink line) which is the usual hidden state h_t. The top path is designed to preserve information because it mainly uses element-wise multiplications and additions controlled by small gates rather than repeated full matrix transformations—this helps avoid shrinking or exploding effects over many steps.

In operation, the model first decides how much of the previous long-term state C_{t-1} to keep (blue "forget" gate),

$$f_t = \sigma(W_f\,[h_{t-1},\, x_t] + b_f)\,,$$

where W_f and b_f are learned parameters, $[h_{t-1}, x_t]$ denotes concatenation (placing the vectors side by side), and σ maps values to $[0, 1]$ like a keep/drop slider. Next, it proposes

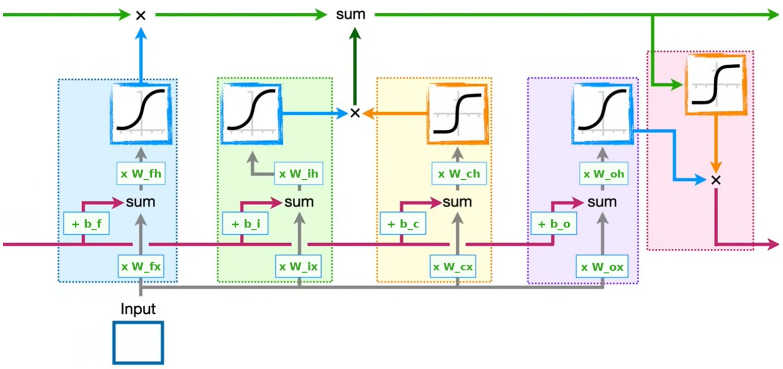


Figure 2.14.: LSTM.
Source: Adapted from Starmer 2025.

new content (yellow) and decides how much to write (green),

$$\hat{C}_t = \tanh(W_c\,[h_{t-1},\, x_t] + b_c)\,, \qquad i_t = \sigma(W_i\,[h_{t-1},\, x_t] + b_i)\,,$$

with tanh producing a candidate in $[-1, 1]$ and $i_t \in [0, 1]$ controlling the write strength; together with f_t this updates the long-term state element-wise,

$$C_t = f_t \odot C_{t-1} + i_t \odot \hat{C}_t,$$

where $\odot$ denotes element-wise multiplication. Finally, the model determines how much of the transformed long-term memory should influence the current short-term state (purple/pink output),

$$o_t = \sigma(W_o\,[h_{t-1},\, x_t] + b_o)\,, \qquad h_t = o_t \odot \tanh(C_t).$$

The result h_t is the updated short-term memory that also serves as the hidden state for the next step.

Combining the idea of sequential processing using RNNs with the stabilization of LSTMs, we thus get the first ANN capable of performing end-to-end, non-rule-based sequence-to-sequence tasks—such as translating from one language to another—where input and output can differ in length, word order matters, and long-range dependencies must be preserved (ibid.). To illustrate this breakthrough, let us shorten the earlier long sentence to “The article was excellent” and see how such a system would produce a Russian translation. The main way to do this is to model the LSTM network using the *encoder–decoder* paradigm as introduced by Sutskever, Vinyals, and Le 2014; Cho et al.

2014. The idea of this architecture is to split the task into two parts, where the *encoder* reads the source sentence and compresses it into a fixed-size summary, and the *decoder* generates the target sentence from that summary (Figure 2.15).

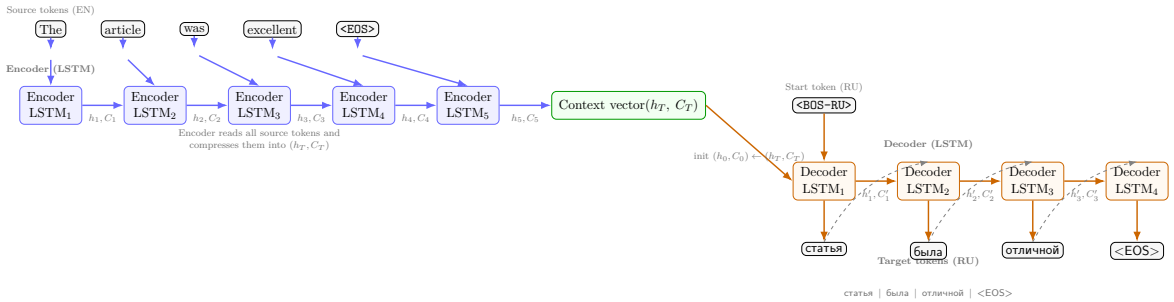


Figure 2.15.: LSTM encoder–decoder.

Source: Adapted from Sutskever, Vinyals, and Le 2014; Cho et al. 2014.

The first step is thus to tokenize the English input as `The | article | was | excellent | <EOS>`.[18] Each token is then mapped to a dense embedding (a learned vector; see Section 2.5) and fed, one by one, into the encoder LSTM (purple), which updates its hidden state h_t and cell state C_t while preserving key cues across the sequence as shown (Figure 2.15). After reading `<EOS>`, the encoder's final states (h_T, C_T) form a summary that compresses the content and order information from the source sentence called the *context vector*.

The decoder (orange) LSTM is initialized with this context vector and a start token `<BOS-RU>`. At each decoding step it takes its current state together with the last produced token and predicts the next Russian token via a probability distribution over the target vocabulary. For our example, it might generate статья | была | отличной | <EOS>: first the subject ("article," feminine singular), then the past-tense copula *была* agreeing in gender/number, and finally the predicate adjective *отличной* in the correct case, stopping at `<EOS>`.

[18] Tokenization splits the input string into a sequence of discrete tokens that the model can process (words, subwords, or characters). In sequence-to-sequence models it is standard to add special symbols such as `<EOS>` (end of sequence) to mark where the input ends, and often `<BOS>` (begin of sequence) to initialize decoding; these markers help the model learn when to stop generating (Manning and Schütze 1999, p. 124).

2.7. The Attention Mechanism

Key terms: information bottleneck, external memory over source positions, alignment, relevance weighting (softmax), step-specific context, dynamic re-access to input, improved long-range dependencies.

Despite the improvements, RNNs (i.e., LSTMs) still have one persistent drawback: a single summary (the final context vector (Figure 2.15)) must carry all information about the source. This makes it difficult for the encoder to compress the sentence (since there are infinitely many possible source sentences and infinitely many possible meanings to those sentences, but they must all be funneled through one fixed-size context vector) and for the decoder to unwrap the sentence (not only because information may be lost in the encoding step, but also since different information may be relevant at different steps and there is just one context reference for the whole sequence) (Bahdanau, Cho, and Bengio 2015). The solution to this problem—the idea that ultimately led to modern LLMs—is to introduce a mechanism called *Attention*.[19] To illustrate what is motivating attention, let us take the translation example from above (Section 2.6) and consider how a human would solve a more complex version of this task. Let us say that the task is not just to translate one sentence but a whole book from English into Russian. Instead of reading the whole English book once, memorizing it, and then working on the Russian translation from that memory—as the RNN-based encoder-decoder model would do it—a human would revisit the original English version many times during the process and use it to figure out how the Russian translation needs to be structured based on the context of the English version. It is this main difference in approaching NLP that is adopted with the attention mechanism summarized in Figure 2.16.

The left part of the figure represents the encoder RNN reading the Russian sentence, and the right part represents the RNN generating the English translation. As we can see, instead of collapsing the entire input into a single vector, the attention mechanism now produces and keeps a sequence of hidden states $(s_1, \ldots, s_m)$ (green columns), where each s_k is the encoder state associated with the source token k. At each time step, the decoder thus sees two things: the current decoder state (red columns) and the sequence of encoder states. For each source position k, the model now determines how relevant this token is at time step t by computing an attention score. If we follow the red lines in Figure 2.16 and zoom in at the blue dot, we can see this part of the network.

[19] Attention was originally introduced in the paper *Neural Machine Translation by Jointly Learning to Align and Translate* from Bahdanau, Cho, and Bengio 2015.

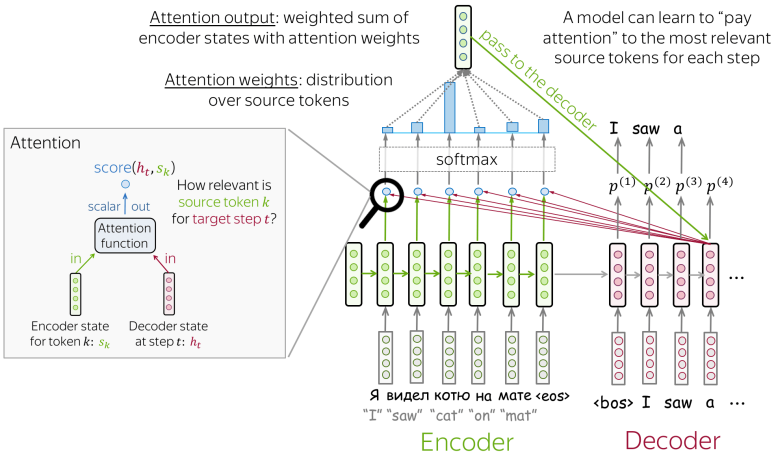


Figure 2.16.: Attention.
Source: Voita 2020.

There are many different ways of computing the attention scores, but generally speaking, the most popular ones are the dot-product, the bilinear function (aka "Luong attention") and the multi-layer perceptron (aka "Bahdanau attention") (see Figure 2.17).

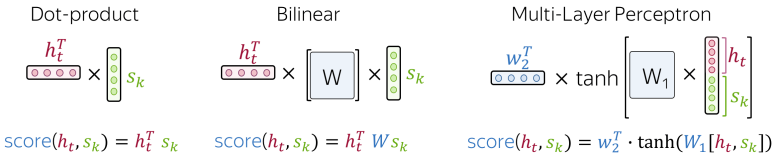


Figure 2.17.: Attention Scores.
Source: Voita 2020.

The resulting score is a single real number that quantifies this relevance: A high value of $e_k^{(t)}$ means that, given the current decoder state h_t, the information stored in s_k is considered very important, and vice versa. These raw scores are then fed into a softmax function, which turns them into a probability distribution over the source tokens, i.e.,

$$a_k^{(t)} = \frac{\exp\left(e_k^{(t)}\right)}{\sum_{i=1}^{m} \exp\left(e_i^{(t)}\right)}, \qquad k = 1, \ldots, m$$

where each $a_k^{(t)}$ lies between 0 and 1, and they all sum up to 1. We can thus interpret $a_k^{(t)}$ as the *attention weight* (Bahdanau, Cho, and Bengio 2015, p. 3) assigned to source token k at decoder step t (illustrated in Figure 2.16 by the blue histogram above the encoder states). Finally, the attention mechanism computes a context vector $c^{(t)}$ as the

weighted sum of encoder states:

$$c^{(t)} = \sum_{k=1}^{m} a_k^{(t)} \, s_k.$$

This $c^{(t)}$ is the *source context* for decoder step t (thus, if a particular encoder state s_k has a large weight $a_k^{(t)}$, its information strongly influences the context). This context sum is then passed to the decoder (Figure 2.16: green line), which updates its state using its previous state, the last generated token, and the newly computed context,

$$h_{t+1} = f(h_t, y_t, c^{(t)}),$$

and produces a distribution over the next output token based on h_{t+1}.

Conceptually, this changes the original encoder-decoder model in two crucial ways (ibid., pp. 5–10). First, there is no longer a single, fixed bottleneck vector: the encoder's entire sequence of states serves as an external memory that the decoder can query repeatedly. Second, the context the decoder uses is no longer the same for all target positions but is dynamically recomputed at each step, allowing the model to "pay attention" to different parts of the input as different words of the translation are generated. This ability to look back at the whole source sequence at every step is what allows attention-based models to handle long and complex sentences much more effectively than the original LSTM encoder-decoder.

2.8. Transformer Networks

Key terms: self-attention only, parallel context construction, query–key similarity, value mixing, scaled dot-product normalization, contextualized representations, architecture template for GPT-style models.

The last step in the evolution that led to state-of-the-art deep learning models such as GPTs consists of what the last term in that abbreviation stands for: the *Transformer*.[20] Originally introduced in the paper *Attention is All You Need* (Vaswani et al. 2017), the Transformer network improves on the previous attention mechanism by eliminating the recurrent part of the architecture and operating solely on attention (since, as the title implies, "that's all you need")—or, to be more precise, on a refined form of attention called *self-attention* (ibid., p. 2).

Although the network structures we have introduced so far (RNNs and LSTMs with or without attention) are in principle capable of handling long sequential inputs, their processing of input sequences—the embedding of words in the appropriate context—is constrained by the RNN property that each time step must wait for the previous one: the model builds context gradually by updating a hidden state one token at a time (see Figure 2.15). Transformer networks, by contrast, leverage self-attention to let every token in a sequence "look at" every other token in the same layer and decide how relevant it is (ibid.). Concretely, with self-attention, each input token is represented in three different ways, corresponding to the roles it can play: it can (1) serve as a *query* and seek information to understand itself better (i.e. ask for information about its surrounding context), it can (2) respond to such requests by acting as a *key*, and it can (3) provide the actual content that will be passed on as a *value*.

Mathematically, this corresponds to three learned linear projections of the same underlying token embedding (Hofmann 2022, 171ff.). If $x_i \in R^d$ is the embedding of the i-th token in the sequence, the model learns three parameter matrices

$$W^Q \in R^{d\times d_k}, \quad W^K \in R^{d\times d_k}, \quad W^V \in R^{d\times d_v},$$

[20] GPT stands for *Generative Pre-trained Transformer*: "generative" because the model produces text token by token, "pre-trained" because it is first trained on a broad next-token prediction objective before any task-specific adaptation, and "Transformer" because it is built from stacked self-attention layers rather than recurrent units (Radford et al. 2018).

and computes

$$q_i = x_i W^Q, \qquad k_i = x_i W^K, \qquad v_i = x_i W^V.$$

Intuitively, q_i encodes what this token is looking for, k_i encodes what this token can offer to others, and v_i packages the information that will actually be mixed into other representations. In other words, the query and key spaces q_i, k_i define a similarity geometry, while the value space v_i carries the payload that will be combined according to these similarities.

To measure how much each key matches each query, the model computes a similarity score for every pair of positions (i, j) via the dot product

$$s_{ij} \;=\; q_i^\top k_j.$$

As in our earlier discussion of word embeddings and cosine similarity (Figure 2.12), this dot product is large when q_i and k_j point in a similar direction and small when they are nearly orthogonal. In matrix form, if we stack all token embeddings row-wise in a matrix $X \in R^{n \times d}$ (for a sequence of length n), we obtain

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V,$$

and the full score matrix $S \in R^{n \times n}$ is given by

$$S = QK^\top, \qquad S_{ij} = q_i^\top k_j.$$

Note that this dot product requires queries and keys to live in the same space (hence $q_i, k_j \in R^{d_k}$); in practice one typically chooses $d_k = d_v$ (and often uses the same dimensionality for queries, keys, and values) for simplicity and efficiency (Vaswani et al. 2017, pp. 3–4). These scores are then normalized row-wise with a softmax to produce attention weights and used to form, for each position i, a weighted sum of the value vectors v_j across the sequence. In compact form this thus gives us self-attention as

$$\text{Attention}(Q, K, V) \;=\; \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V,$$

where the softmax converts each row of $QK^\top$ into a probability distribution over keys and the factor $\sqrt{d_k}$ stabilizes the scale of the dot products. The result is that each token's new representation becomes a context-sensitive mixture of value vectors from all positions, weighted by how well their keys match its query (ibid.).

2.9. Summary

In this technical introduction, we reconstructed the main steps in the development from early, rule-based AI to modern neural language processing, in order to establish the minimal conceptual background needed for the CLERE analysis that follows.

We began with *symbolic AI* (Section 2.1) by framing a classical rationalist ideal of cognition (Descartes' method) as an ideal of explicit, inspectable, stepwise reasoning. Against this background, we traced how advances in formal logic (Boolean algebra; Fregean predicate logic with quantifiers) made it plausible to treat thinking as formal symbol manipulation, and how this perspective was operationalized in early AI through programs like the Logic Theorist and the General Problem Solver. The GPS example highlighted the state-goal-operator framework and means–ends analysis, culminating in the physical symbol system hypothesis as a paradigmatic statement of the symbolic program.

We then turned to *subsymbolic AI* (Section 2.2) by shifting the modeling inspiration from explicit rules to biological nervous systems. Starting from threshold-style neuron models, we introduced Hebbian learning as the decisive conceptual step toward adaptive, parameterized networks, and from there the perceptron (Section 2.3) as a first operational supervised learner, along with its separability limits illustrated by XOR.

This motivated the move to *multilayer networks* (Section 2.4). We introduced hidden layers as learned representations and explained how deep learning relies on differentiable loss functions and backpropagation for credit assignment. This provided the basis for non-linear decision-making and for framing universality as a representational—not automatically practical—guarantee.

Finally, we connected deep learning to *natural language processing* (Section 2.5) by outlining why language demands specialized representations and architectures. We moved from the limitations of one-hot encodings to distributional word embeddings, then to sequence models (Section 2.6) and the encoder–decoder paradigm, introduced attention (Section 2.7) as a remedy to the fixed-context bottleneck, and culminated in the Transformer (Section 2.8) as the current standard architecture built around self-attention.

With this trajectory in place—from symbolic rule systems to learned distributed representations and attention-based sequence modeling—we can now turn to the CLERE section, where we distill these technical components into a set of explicit modeling assumptions about learning, representation, and performance that will serve as the stable reference point for the epistemic analysis developed in Chapter 3.

2.10. CLERE—Conceptual Learner for Epistemic Reasoning Experiments

In this last section, we present CLERE (Conceptual Learner for Epistemic Reasoning Experiments) as our hypothetical artificial neural network that serves as the object of analysis for the remainder of this thesis. The purpose of CLERE is twofold: (1) it consolidates the machine-learning theory from the preceding chapter into a set of explicit modeling assumptions about learning, representation, and performance, so that we can conduct the discussion with a concrete point of reference, and (2) it provides a bridge from the technical to the philosophical by abstracting epistemically relevant concepts from their technological embedding.[21]

Training Data The primary object from which CLERE learns is its *training data*. As a first step, it is therefore important to note that, although the term "data" derives from the Latin *datum* ("[a thing] given") (Merriam-Webster n.d.),[22] the way it is used in the context of CLERE in no way refers to something simply given, but rather to the outcome of multiple prior choices about what to measure, how to sample, and how to clean, format, and label records (Gitelman 2013; Kitchin and Lauriault 2014). From today's vantage point, it is therefore easy to forget that we have only recently transitioned from a world in which data were a scarce resource to one in which they are abundant.[23] A major part of *what* CLERE comes to "know" is thus already constrained by these upstream design decisions about what is measured, collected, included, and excluded.

[21] A step that cannot be taken for granted since, as we have seen (Chapter 1), much of the debate relies on a shared vocabulary whose terms do not carry the same meaning across disciplines. Expressions such as "learning," "representation," "meaning," "information," or even "knowledge" are often used in machine learning as operational labels for measurable relations between data, parameters, and performance, whereas in philosophy they function as normatively loaded notions tied to justification, understanding, truth, and agency. If these shifts in sense are not made explicit, the discussion easily slides into equivocations and category mistakes: one may read epistemic significance into what is merely a technical success criterion, or conversely criticize a model for failing to satisfy philosophical conditions that were never part of its operational specification. CLERE is therefore introduced precisely to keep these levels apart while still allowing us to translate between them in a controlled way.

[22] More precisely, the neuter past participle of *dare* ("to give"); see (Merriam-Webster n.d.).

[23] This historical shift is visible in NLP resources: in the early 1990s, flagship corpora such as the Penn Treebank comprised on the order of a few million words and required extensive manual annotation (Marcus, Santorini, and Marcinkiewicz 1993), whereas web-scale pipelines today can ingest billions of pages per crawl (e.g., Common Crawl's December 2024 release reports about 2.64 billion pages collected from 1–15 December 2024) (Nagel and Vaughan 2024).

Performance Improvement The learning procedure itself is operationally defined by an improvement in task performance through what is generally denoted as "experience." Tom Mitchell's classic definition states that a system like CLERE "[...] is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell 1997, p. 2). The term "experience" thus references the training exposure that provides a system with information about how well it is doing. For CLERE, experience E is the cumulative history of training steps: mini-batches sampled from the training corpus, forward passes that produce predictions, and backward passes that compute an error signal and adjust parameters accordingly.

Error-Driven Self-Modification Unlike symbolic AI systems that are programmed with explicit rules, CLERE modifies itself by tuning the strengths of connections (weights) in response to data. Each training example provides a loss (error signal) indicating how far the output is from the target. This loss is then propagated through the network and used to make small, local changes to parameters that, in aggregate, move CLERE toward better performance under the selected objective. What is considered a "useful feature" to CLERE is thus determined entirely by what helps to reduce the loss in that training regime.[24] Over many such examples, this process embeds statistical regularities of the data into the network's parameters. The resulting internal "feature sensitivities" are thus not hand-coded or fixed in advance as semantically meaningful components; they are an emergent by-product of error-driven adjustments of connection strengths under the chosen objective.

Distributed Internal Representations The "knowledge" that CLERE accumulates is stored in a distributed fashion across many neurons and layers, rather than in any single unit or explicit symbol. In connectionist terms, each concept or feature is represented by a pattern of activations spread over numerous units, and conversely each unit participates

[24] An important point, because it means that the system's internal organization is relative to the task specification and the training signal, and not a direct mirror of the world "as it is." In practice, this is visible in so-called *shortcut learning* and *dataset bias* (Geirhos et al. 2020): if a spurious correlate reliably predicts the label in the training distribution, gradient-based training will tend to exploit it. A standard illustration is an image classifier trained to distinguish *wolves* from *huskies* on a dataset in which wolf pictures disproportionately contain snowy backgrounds (Ribeiro, Singh, and Guestrin 2016, p. 8); the model may internalize "snow" as a highly predictive feature and therefore misclassify a husky in snow as a wolf, or a wolf on grass as a husky. Although the learned representation is successful relative to the loss on the training regime, it does not necessarily track the properties that humans take to be constitutive of the category.

in representing many different concepts. For example, there is no single neuron that exclusively stores the concept "cat"; instead, the concept "cat" corresponds to a particular activation pattern across a subset of CLERE's units (which might also be involved in representing related animals, fur textures, etc.). This distributed encoding enables generalization and fault tolerance (since "knowledge" is not wiped out by the loss of one unit) but also means we cannot point to a specific weight or node as "the cat memory." "Knowledge" in CLERE is thus implicit in the overall configuration of weights.

Learning as Compression One of the most important aspects of CLERE's learning ability is to separate important information from unimportant information (noise). Just like the human brain optimizes cognition by minimizing redundancy and compressing sensory input, retaining only what is informative,[25] CLERE's deep learning network builds an internal bottleneck representation that encodes just the relevant features of the input for predicting the output, effectively forgetting or filtering out irrelevant details. According to Naftali Tishby's information bottleneck theory (Tishby, Pereira, and Bialek 2000), the adequate description of CLERE's learning strategy is therefore "[...] not remembering the information per se but forgetting unimportant elements of it" (Mugleston et al. 2025, p. 2).

High-Dimensional Encoding The way that what we call CLERE's "knowledge" (for now) is encoded is best understood geometrically: learning yields high-dimensional vector spaces in which tokens, phrases, and more abstract patterns are represented as points and directions. During training, CLERE learns a representation in which downstream tasks become linearly tractable—i.e., the "messy" structure of the raw input is "unfolded" into a space where simple separators (hyperplanes) can implement complex distinctions.[26] High dimensionality is helpful here because it provides many degrees of freedom for distributed coding: different aspects of similarity (syntax, topic, register, entailment cues,

[25] This is the core idea of the *efficient coding* (or redundancy-reduction) hypothesis (Barlow 1961, pp. 217–234): sensory systems are hypothesized to allocate representational resources so as to reduce statistical redundancy in the input while preserving information that matters for downstream use. For a modern overview of efficient coding in terms of natural image statistics and neural representation, see Simoncelli and Olshausen (2001).

[26] This is the core intuition behind classical *kernel methods*: a (possibly implicit) feature map $\phi(x)$ embeds data into a higher-dimensional space in which linear decision boundaries correspond to non-linear boundaries in the original space (the *kernel trick*). Support Vector Machines are the canonical example (Cortes and Vapnik 1995; Schölkopf and Smola 2002). Deep networks differ mainly in that they *learn* such feature mappings from data (rather than fixing ϕ in advance), yielding task-shaped representations.

etc.) can be captured in partially independent directions, enabling robust generalization beyond memorized instances.[27]

Conclusion With regard to the next chapter, we thus conclude the discussion with the following observations: (1) whatever we are prepared to call CLERE's "knowledge" is to a significant extent inaccessible to direct inspection: even if the trained network can reliably deploy what it has learned across a variety of tasks, there is no straightforward way to "look inside" the model and recover that knowledge as an explicit set of reasons, rules, or propositions. It is stored primarily as numerical parameter values and distributed activation patterns, and can only be investigated indirectly—through empirical probing, interpretability techniques, and controlled interventions—rather than by simply reading off a theory from the system's internal states. (2) As a consequence, CLERE seems, prima facie, to occupy an epistemically unusual position: it is neither a classical symbolic AI system, nor a conscious biological agent. Its internal representations are subpersonal in the strict sense—i.e. they are not sentences in an inner language and they are not available to the system as articulated reasons—and yet, it demonstrates strikingly flexible, context-sensitive performance that exceeds, in many ways, the capabilities of its symbolic predecessor as well as its human creators.

[27] This representational gain, as we shall see in more detail later on, has an important downside: In high dimensions, intuitive notions like "closeness" become unstable (distances can concentrate and neighborhoods become counterintuitive), so similarity is no longer a transparent, human-readable relation but a task-shaped metric induced by training (Beyer et al. 1999; Aggarwal, Hinneburg, and Keim 2001). This contributes to the black-box character of CLERE: small, perceptually irrelevant input changes can push an example across a decision boundary in representation space and flip the output, not because "the world" changed, but because CLERE's learned geometry partitions the space in ways that do not align with human perceptual categories.

3. Epistemic Frameworks

As we have seen in the previous chapter, there are two main paradigms within the intellectual and technical evolution(s) that eventually led to the possibility of machine learning systems such as CLERE: symbolic (Section 2.1) and sub-symbolic (Section 2.2) AI. Applied to the question of epistemology, this means distinguishing between a *cognitive* (epistemic) and a *connectivist* information paradigm (Gramelsberger and Sigwart 2025, p. 262): the former corresponds to the view that intelligent performance is, at its core, a matter of operating on explicit, structured representations according to rules, and the latter corresponds to the view that competence is grounded in learned connectivity. For much of the history of thinking about the epistemic implications of ostensibly "intelligent" machines, the focus of the analysis was set on the former: systems based on formal, rule-based symbol manipulation. Accordingly, arguments in favor of whether such artificial entities, with their increasingly improving problem-solving capabilities, should be considered epistemic in the way that we consider human problem solving to be epistemic were met with significant pushback. Philosophers like John Searle have taken the influential position that no matter how complex these symbol-manipulating machines may become, in the end, their achievements can be broken down to an endless chain of purely syntactic relations among symbols. The thought experiment underpinning this idea is Searle's *Chinese Room* (Searle 1980): imagine a person who knows no Chinese locked in a room with a rulebook that specifies, step by step, how to transform incoming Chinese character strings into new strings that native speakers outside judge to be appropriate replies. From the outside, the exchange is indistinguishable from that of a competent speaker; inside, however, the person is merely applying syntactic procedures with no grasp of semantic content. Searle concludes that neither the person nor the room thereby understands Chinese, and, by analogy, that a computer executing a program likewise performs formal symbol manipulation without genuine understanding—it is just "symbols in, symbols out" (ibid.).

However, in light of the major advances within the sub-symbolic, connectivist paradigm, the question arises: *what if we put CLERE in the Chinese room?* What if, instead of a rulebook operator, there is someone in the room who has been shown numerous examples, counterexamples, and corrective feedback, and thus learned to form associations between them and eventually built an internal representation that allows for competent replies? In what follows, we cover the philosophical theory needed to ground such a question. First, we determine how the term "knowledge"—or, more precisely, "propositional knowledge"—

can be defined within the context of machine learning. The core question here is: what does it mean when we say that "x knows that p" where x is an entity—such as CLERE—whose knowledge is in question, and p is a declarative statement such as "the word for 'philosophy' in Chinese." Second, we introduce the notion of tacit knowledge in order to capture forms of "knowing" that are not straightforwardly expressible as declarative propositions. The guiding question here is what it means to attribute knowing how to a system like CLERE: how claims like "CLERE knows how to reply to my question in perfect Chinese" relate to, but also differ from, claims of the form "CLERE knows that p."

3.1. Propositional Knowledge

3.1.1. Justified True Belief

Key terms: truth-belief-justification, rationalism, empiricism, a priori, a posteriori, Hume, causality, induction, Kant, conditions of experience.

Within traditional epistemology, one of the core foundations on which much of the debate around the questions of what propositional knowledge is, how it is formed, and when it occurs is centered around the notion of *justified true belief* (JTB). According to JTB—which first appeared in Plato's *Theaetetus*, where Socrates is confronted with the proposal that knowledge is "true belief with an account"—i.e., *logos*—the three conditions necessary for knowledge are (1) truth, (2) belief, and (3) justification (Ichikawa and Steup 2024). JTB thus yields the definition:

S knows that p iff: (1) p is true, (2) S believes that p, and (3) this belief is justified. (JTB)

The first condition states that knowledge must in some way or form imply truth. One cannot know something that is not true, even if there is a corresponding belief and a justification of that belief. According to JTB, it does not make any sense to say that S knows that the capital of Switzerland is Zürich, since the proposition contained within that statement is false. Second, JTB requires the subject to accept the proposition in question, i.e., to have some sort of (coherent) belief in it. According to Sartwell 1991, this implies that: "I cannot have the belief that Goldbach's conjecture is true and fail to have any related beliefs. The belief is constituted as a belief within a system of beliefs" (ibid., p. 158). Thus, statements such as (correct) guesses are excluded from knowledge since they do not constitute a coherent system of belief. Third, JTB requires that knowledge be based upon justification so that even if true, beliefs formed through erroneous reasoning do not qualify as knowledge.

It becomes apparent that, in particular, the third condition of JTB raises the difficult question of what reasons suffice to justify a belief. This issue (re-)divided the epistemological debate of the modern era along the already existing lines of the *problem of universals*.[1] The two opposing positions are generally grouped into the rationalist and the empiricist schools of thought. We have already introduced one of the most important advocates of the former in the context of the intellectual roots of symbolic AI. René Descartes's attempt to derive a systematic method by which to avoid contradictions in the sciences and generate well-founded knowledge is based

[1] That is, the ancient question of whether properties of an object, such as color and shape, exist within or beyond that object. See (Rodriguez-Pereyra 2000) for an overview.

upon the argument that there are three types of "ideas" that underlie knowledge: (a) ideas that are based upon perception, (b) ideas that are generated by imagination, and (c) ideas that are innate ("ideae innatae") (Gramelsberger and Sigwart 2025, p. 45). The difference to Plato's theory of forms is that Descartes's *ideae innatae* are not conceived as independent entities but rather as a result of systematic reason: they are clear and evident and thus form the a priori basis for certainty of knowledge. The main aspect of rationalism that is relevant to our question is therefore the notion that (at least a portion of) our justified knowledge rests on structural grounds.[2]

In contrast to rationalism, the empiricist school of thought rejected the notion of innate ideas and, by extension, the rationalist approach of obtaining them. Empiricism defines knowledge as a result of what can be observed, recorded, and generalized from sensory impressions and experience. Hume's work in particular was strongly influenced by the developments in neurological research of the 18th century that quickly led natural scientists to the materialist assumption that the concept of mind or spirit can be reduced to stimuli in the nervous system. In this context, Hume wanted to demonstrate that nerve stimuli and the perceptions based on them are indeed the basis of human action and thought, but that human nature—including the concept of knowledge—is not exhausted by them (Gramelsberger and Sigwart 2025, p. 52). With his work in *An Enquiry concerning Human Understanding* (1748) and *A Treatise of Human Nature* (1739), he thus formulated two important problems with respect to the question of knowledge: the *problem of causality* and the *problem of induction*.

According to Hume, true causality is a process that presupposes that (1) the cause precedes the effect (temporal priority); (2) cause and effect occur next to—or at least continuously with—one another in space and time (spatiotemporal contiguity); and (3) there is an objective tie that carries the effect given the cause (necessary connection) (Hume 1978, pp. 75–77). The argument that Hume puts forward with the problem of causality is that we, as human observers, never perceive this necessity in the objects or events themselves. All we actually see is one event followed by another—what he calls *constant conjunction*: events of type A are regularly followed by events of type B, and from these repeated pairings, habit produces a natural expectation that B will follow A. Therefore, Hume argues that the idea of necessity does not come from reasoning but from the mind's internal feeling of being determined—after repeated experience—to expect the effect upon the appearance of the cause.

Similarly, he argues with the problem of induction that there is no rational argument that

[2] Leibniz, who, as we have seen (Section 2.1), advanced both epistemology and mathematics, formulated the *principle of sufficient reason* (PSR), according to which "[...] no fact can be real or existent, no statement true, unless there be a sufficient reason why it is thus, and not otherwise" (Leibniz 1902, p. 258). For an overview of the PSR and its role in Leibniz's rationalism, see (Melamed and M. Lin 2023).

could justify the claim that regular experience will continue to occur in the future, since there is no way to prove the uniformity of nature without circularity. Take Bertrand Russell's famous example illustrating this problem:

> Domestic animals expect food when they see the person who usually feeds them. We know that all these rather crude expectations of uniformity are liable to be misleading. The man who has fed the chicken every day throughout its life at last wrings its neck instead, showing that more refined views as to the uniformity of nature would have been useful to the chicken (B. Russell 1912, pp. 97–98).

Hume thus leaves us with two notions that fundamentally undermine the rationalist's claim to justified propositional knowledge: (1) experience shows only constant conjunction, never a necessary connection (causality) in the objects; (2) our belief that the future will resemble the past (induction) rests on custom (habit), not reason.

Building on the work of Hume, Kant splits the difference between the rationalist and the empiricist positions by distinguishing between *a priori* and *a posteriori* forms of knowledge, in which the former is independent of experience, like mathematical truths, and the latter relies on empirical evidence, such as sensory observations (Mugleston et al. 2025, p. 2). According to Kant, although it is evident that all our knowledge begins with experience—"[f]or how is it possible that the faculty of cognition should be awakened into exercise otherwise than by means of objects which affect our senses" (Kant 2018, p. 17)—this does not imply that "all arises out of experience," since experience itself is already organized by a priori contributions of our mind (ibid., p. 17). Although there is no straightforward way to untangle the complex structure of this line of argumentation, a broad outline can be reconstructed as follows: according to Kant, space and time are forms of intuition (Formen der Anschauung) that structure sensible givenness (Sinnlichkeit) into appearances (Erscheinungen), while the intellect (Verstand) supplies categories and—through schematism—applies them to appearances (Gardner 1999, 66ff.). Thus, propositions, for Kant, are acts of judgment that bring intuitions under concepts according to rules; when such assent (Fürwahrhalten) rests on grounds of cognition that are subjectively sufficient (conviction) and objectively sufficient (certainty) for any rational subject, the result counts as knowledge (Wissen) (ibid., 66ff.).

Summary The historical debate from Plato through Descartes, Hume, and Kant reveals a variety of ways of understanding the justificatory element in knowledge: for the rationalist, justification ultimately rests on a priori structures of reason; for the empiricist, it is grounded in sensory impressions and their habitual associations; for Kant, it arises from the synthesis of a priori forms and empirical content. Despite these differences, a common core emerges that is crucial for our purposes: propositional knowledge is at least a matter of a subject holding a

true belief under some normatively constrained form of support that distinguishes mere opinion from cognition with a claim to objectivity. In what follows, this notion of JTB will serve as the background against which Gettier's challenge is formulated and against which more fine-grained internalist and externalist accounts of justification, relevant to the epistemic status of systems like CLERE, will be assessed.

3.1.2. Gettier and the Reframing of Justification

Key terms: Gettier problem, internalism, externalism, propositional justification, doxastic justification, supervenience thesis, access internalism, mentalism, Bayesian credences, truth-conductiveness, process reliabilism.

The notion of JTB has been most famously challenged by Edmund Gettier, who proposed that truth, belief, and justification are necessary but not sufficient conditions for knowledge, since one can possess a justified belief that is true and nevertheless fail to know because the reasoning includes a mistake—a *false lemma*. To illustrate this line of argumentation, let us take the following example: let us say that Mira is birdwatching in a city park. She hears what sounds exactly like a nightingale and, relying on her well-trained ear and prior experience, forms the belief that there is a nightingale in the park right now. Unknown to her, the sound comes from a nearby phone playing a nightingale call. However, behind a dense shrub on the other side of the path, an actual nightingale is present at that very moment. Mira's belief is true and well-justified by her evidence; but the justification hinges on a false assumption (that the sound was produced by a nightingale) (K. R. Harris 2024, 986ff.). Hence, she has a justified true belief but, as Gettier would argue, no knowledge.

Gettier-style cases thus show that having good reasons that cohere with one's evidence can still yield a true belief by luck. This pushes the debate about propositional knowledge toward two main families of responses. The first, typically labeled *internalism*, holds that since bare JTB is too permissive, the grounds that make a belief justified must be, in an important sense, "inside" the subject—available to reflection or at least fixed by the subject's mental life. On this view, justification is a matter of one's reasons/evidence, and knowledge is secured by strengthening the subject's epistemic position. The second, *externalist*, family argues that justification must be appropriately connected to the world. On this view, knowledge depends on factors "outside" of the subject's perspective—e.g., whether the belief was formed by a reliable process, is safe from error, or results from properly functioning cognitive faculties in a suitable environment.

3.1.2.1. Internalism

Internalist responses to Gettier start from the thought that what turns a mere true belief into a justified one must be fixed "from within" the subject's epistemic perspective. It is important to note that within this debate, the term "justification" is connected to two different questions: First, there is the question whether the subject has good epistemic support for a proposition p at all. This is typically called *propositional justification* (Silva and Oliveira 2023): S has propositional justification for p when S's evidence (or reasons) supports p to the required degree. Second, there is the question whether S actually believes p on the basis of that support, rather than for some epistemically irrelevant reason. This is called *doxastic justification* (ibid.). In the standard picture, doxastic justification requires propositional justification plus an appropriate *basing relation*,[3] between the subject's reasons/evidence and the belief. In what follows, we focus primarily on propositional justification, since internalism is most plausibly understood as a thesis about what determines propositional justification (rather than a full account of the basing relation) (Korcz 2021).

The core claim of internalism is often formulated as a *supervenience thesis* (Littlejohn 2025): if two subjects are alike with respect to everything that constitutes their epistemic point of view, then they must be alike with respect to what they are propositionally justified in believing. Justification, on this view, is meant to be a norm of rationality. It evaluates how well a subject's doxastic attitudes respond to what is available from the subject's point of view, rather than to facts that may obtain entirely outside of it. A popular example used to strengthen this line of argumentation is the so-called "new evil demon problem" (ibid.)—or, what might be called "the matrix dilemma": imagine an all-powerful, malicious Cartesian demon (or an evil sentient robot) that systematically deceives subjects about the external world. From the inside, the victims of the demon have experiences, memories, and inferential dispositions that are, ex hypothesi, indistinguishable from subjects outside the demon's reach. They seem to see a computer in front of them, to remember having walked to the library this morning, to feel the keys under their fingers, and so on. In reality, however, there is no computer, no library, no physical world of the familiar sort at all; all of their experiences are perfectly orchestrated illusions produced by the demon.

According to the supervenience thesis, such subjects remain just as rational as we are in taking things to be the way they appear (ibid.). If their total experiential situation and internal perspective are a perfect match to ours, then—so the internalist argument goes—they are equally justified in believing that there is a computer in front of them, that they once walked to the library, and that there are hands at the end of their arms. After all, the mere possibility

[3]For the basing relation and its role in distinguishing propositional from doxastic justification, see (Korcz 2021).

that we as well might be radically deceived by such a demon (i.e., live in the matrix) does not undermine the justification of our ordinary beliefs, so why should the actual deception of the demon's victim do so in their case?

Following this line of argumentation, the question then arises: what exactly are the "internal states" that the subject's perspective is supposed to range over? Without specification, there are several candidate notions that come to mind: an inner state might be understood in bodily terms, such as the subject's heart rate or body temperature (bodily states); in neural terms, such as a specific pattern of firing in the visual cortex (brain states); or in psychological terms, such as the subject's current visual experience of holding a cup of coffee, together with the accompanying seemings and attitudes (mental states) (Littlejohn 2025). It is this question that divides the internalist framework into two main families.

Access Internalism On the one hand, access internalism takes the position that, no matter what the underlying state, the requirement for a justified belief must be—at least in principle—accessible to the subject by reflection (ibid.). On this view, a belief is justified only if the subject has (or could have, given sufficient time and attention) reasons that they can bring into view and assess as counting in their favor. The relevant "justification-determining factors" (J-factors) (Poston 2026) are thus those that can figure within the subject's deliberative standpoint: what their experience is like, what they remember, what they take themselves to have perceived, what they already believe, and what they can recognize as logical or evidential relations among these states (for instance, that two of their beliefs conflict, or that a piece of evidence supports one hypothesis better than another).

If we were to apply this notion to Mira, who is birdwatching in the park (Section 3.1.2), it becomes evident that access internalism delivers the intuitive verdict that she is propositionally justified. From her point of view, she hears what sounds exactly like a nightingale, she remembers having successfully identified nightingale calls in the past, she believes that she is currently in an environment where nightingales are present at this time of year, and she has no defeating information to the contrary. All of these factors are, in principle, available to her upon reflection: if asked why she thinks there is a nightingale in the park, she could cite precisely such experiences, memories, and background beliefs—and, since further facts, such as that there is a speaker nearby producing the sound, are not reflectively accessible to her, they do not form part of her justificatory basis.

Mentalism The problem with access internalism, and this is the main argument that the second internalist family brings forward, is that the notion of access itself can be read as a "thinly disguised epistemic term" (ibid.) that results in a circularity.

> To have access to some fact is just to know whether or not that fact obtains. This is problematic for an accessibilist because he analyzes justification in terms of access and then uses the notion of justification to partially explicate knowledge. In short, if 'access' is an epistemic term then any analysis of knowledge that rests upon facts about access will be circular (ibid.).

What determines justification according to the mentalist view is thus not access, but the subject's mental states themselves. On this approach, often associated with Feldman and Conee (2001), internalism is understood as the thesis that a subject's justificatory status supervenes entirely on how things are mentally for that person: on their occurrent and standing beliefs, perceptual experiences, seemings, memories, and so on, irrespective of whether all of these are presently available to reflection or not (ibid.).

One of the advantages of mentalism, so Feldman and Conee argue, is that it preserves a clear internalist supervenience thesis without relying on the potentially circular notion of access (Poston 2026). Justification is determined by which mental states the subject in fact has, not by which states it can currently bring to consciousness or know that it is in. However, mentalism does face the objection that, by dropping any requirement of reflective availability, it risks diluting what is distinctive about internalism: if some of the justificatory basis may be constituted by mental states that are in principle beyond the subject's reflective reach, critics such as Bergmann (2006) have questioned whether the resulting view still earns the label "internalist" in the traditional sense.

What matters for our purpose, however, are not so much the details of the differences between accessibilism and mentalism but the fact that both families treat justification as fixed by an internal profile—whether formulated in terms of accessible reasons or more broadly in terms of mental states—and that both leave open the further question of how, exactly, these internal states should be regimented and evaluated. With that question in mind, we turn to an internalist framework that offers a precise norm for how such a profile should fit together and evolve under incoming evidence—an account that will later help us in addressing the question of whether and in what sense CLERE can be said to satisfy the internalist demands of rationality.

Bayes Credences Within this general internalist landscape, Bayesian epistemology can be understood as a more specific and formally articulated framework for regimenting a subject's epistemic perspective. According to the Bayesian approach—tracing back to Thomas Bayes who, next to his famous contributions to the field of statistics,[4] is also often read as a pioneer of internalist epistemology (Otsuka 2022, 53ff.)—a belief is not, as we have treated it so far,

[4]Bayes' theorem provides one of the fundamental building blocks for modern probability theory and many machine-learning methods; see (Murphy 2012).

an all-or-nothing affair: the question is not so much whether we believe that p or do not believe that p, but rather how strongly we believe that p. Beliefs, so Bayes argues, can come in different strengths; what he calls "degrees of belief" or "*credences*" (H. Lin 2024). For example, consider the proposition "We exist." As followers of Descartes, we are convinced that even if an all-powerful demon tried to deceive us about everything else, the fact that we exist would remain unchanged ("cogito ergo sum"); therefore, we strongly believe that we exist. But since we are also interested—albeit skeptical—readers of the many-worlds interpretation of quantum theory,[5] we are only fairly confident that we exist right here and right now. Furthermore, we strongly doubt that we will exist in 70 years (even given current longevity trends), and we are very confident that we will not exist in 100 years.

In Bayesian terms, we thus assign to any proposition A a probability $P(A)$ that represents our degree of confidence—our credence—in A. Bayes then imposes two governing norms on such credences and their evolution under evidence (H. Lin 2024): First, in order to guarantee synchronic coherence (*probabilism*), the credence function $P\colon \mathcal{F} \to [0,1]$ must satisfy the probability axioms of non-negativity, additivity, and normalization (total mass 1). This provides a decision-theoretic rationale for the axioms as norms of rational belief—i.e., avoids incoherences that lead to contradictions such as Dutch book style arguments.[6] Second, in order to guarantee diachronic coherence (*conditionalization*), the updates over time must match the earlier conditional probabilities given the new evidence (H. Lin 2024). Hence, if we have a proposition A (which can be true or false) and new information or evidence B (which can also be true or false), Bayes shows that our posterior belief $P(A|B)$ is calculated by multiplying our prior belief $P(A)$ by the likelihood $P(B|A)$, or the probability of evidence B given that the proposition A is true. The conditional probability of A given that B is true is thus:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}, \qquad P(B) \neq 0. \tag{Bayes' theorem}$$

Whereas within the previous debate on JTB (Section 3.1), justification was inherently connected to providing time-independent reasons—i.e., proofs, observations, reliable testimony, explanatory considerations—that are taken to support p independently of the subject's evidential perspective

[5] The many-worlds (Everett) interpretation holds that, in situations (famously in quantum measurements) where more than one outcome is possible, reality does not "select" just one; rather, all outcomes are realized in different, mutually independent branches; see (Wallace 2012).

[6] Imagine we allow for credences that violate an axiom of probability. Let us say $P(A) = 0.6$ and $P(\neg A) = 0.5$. Let us further say those credences represent bets that someone offered us a 1-euro payout ticket at our prices (i.e., we buy A for 0.6 euro and $\neg A$ for 0.5 euro, which amounts to a total of 1.10 euro). In any case, since it is always either A or $\neg A$, we receive 1 euro, which amounts to a sure loss of 0.10 euro for every scenario. Arguments like these are called Dutch book arguments (H. Lin 2024), which illustrate that if our credences (or our plan for updating them) violate the probability axioms (or conditionalization), then a bookmaker can construct a finite portfolio of bets, priced at our own odds, that guarantees us a loss in every possible state.

and reasoning policy, Bayes argues that S is justified to believe p iff the process is based on a coherent set of credences and a rule-governed transformation from old credences to new ones when evidence arrives—that is, as we have just seen, the credence function is probabilistically coherent and the transitions from prior to posterior are governed by conditionalization.[7] This shifts the focus from a static ledger of reasons to procedural rationality under (subjective) evidence—a shift that can, as we shall see in the next Chapter 4, in principle be argued to allow for non-human agents to hold justified true belief—i.e., credence.

Summary The internalist frameworks we have covered—from accessibilism and mentalism to the Bayesian picture of credences—agree that justification is fixed by a subject's internal epistemic profile, understood in terms of experiences, mental states, or a coherent credence function. While this provides an account of what it is for a subject like Mira to be rational from her own point of view, it does not resolve the Gettier problem: as long as justification is purely internal, a belief can be both justified and true while still owing its truth to epistemic luck. This observation motivates the externalist turn to reliabilist accounts of justification which aim, as we shall see in the following section, to supplement internal rationality with an appropriate truth-connection between belief-forming processes and the world.

3.1.2.2. Externalism

Externalist responses to Gettier begin from a shift in what the justificatory condition is supposed to accomplish. Internalist accounts, in their different ways, capture an important normative idea: a subject can be epistemically responsible (rational) given what is available from their point of view. Yet Gettier-style cases suggest that responsibility, coherence, and reflective defensibility do not by themselves secure what knowledge is centrally supposed to exclude: true belief that is correct merely by luck. Externalism takes this diagnosis seriously and treats justification (or whatever plays its role in an analysis of knowledge) as answerable not only to the subject's perspective but to the world in which the belief is formed. The guiding thought is that epistemic norms are not exhausted by internal rationality; they are also meant to discriminate beliefs that are "well-connected" to their truth-makers from beliefs whose truth is accidental or otherwise flawed in an epistemic sense (Littlejohn 2025).

An important factor within this line of thought is the concept of truth. Even though most post-Gettier approaches to propositional knowledge no longer center around the search for an infallible criterion of truth—as was the case with traditional epistemologists such as Descartes—the notion of justification is still expected to be in some shape or form *truth-conductive*:

[7] What constitutes knowledge under these norms is then typically defined according to a context-dependent threshold where $P(A|B) \geq t$ (Otsuka 2022, 87ff.).

meaning that it must serve as a guide toward the truth, "[...] perhaps not infallible but nonetheless reliable to a certain degree" (Otsuka 2022, p. 58). On that view, the externalist critique of internalism is that if one is not willing to abandon the idea that the term "truth" is somehow connected to the world external to the subject holding a belief—a commitment that all externalist notions presuppose (Littlejohn 2025)—it remains unclear how the justification of a belief in terms of its coherent relationship with the subject's other beliefs should contribute to its truth-conductiveness.

The externalist appeal to truth-conductiveness thus reframes the internalist notion of likelihood: the objective likelihood of a belief given a body of evidence is no longer tied primarily to a disciplined learning procedure or a coherent system of credences—as in the Bayesian picture—but to the strength of the correlation between the truth of the belief and the body of evidence in the actual world. Consider the example of the litmus test (Poston 2026). If one applies a liquid to litmus paper and it turns red, then the objective likelihood that the liquid is acidic is very high; that is, in the relevant range of nearby worlds, red litmus reliably co-occurs with acidity. Yet the existence and robustness of this correlation are not themselves reflectively accessible to an ordinary subject; they depend on underlying chemical regularities, the manufacturing process of the litmus paper, and so on. A subject may therefore be objectively well-placed with respect to the truth—their belief that the liquid is acidic is formed in a way that tends to produce true beliefs—without having any internal grasp of why the indicator is reliable (ibid.).

Process Reliabilism One of the most important concepts that develops the externalist conception of JTB into a systematic epistemic framework is Goldman's process reliabilism (PR) (A. Goldman and Beddor 2021). The guiding idea of Goldman is simple: a belief is epistemically justified, in the relevant sense, when it is produced and sustained by a cognitive method that reliably results in truth-conductive propositions.

> **Process Reliabilism.** If S's believing p at t results from a reliable cognitive belief-forming process (or set of processes), then S's belief that p at t is justified (ibid.).

The first thing to notice about this definition is that, in contrast to internalism, the main emphasis of the term "justification" is no longer propositional but doxastic. Goldman is not primarily interested in the subject's merely having adequate evidence for p in a reflective sense, but in whether a particular token belief that p is epistemically acceptable given the way in which it was produced (ibid.). Second, Goldman introduces a new variable that did not figure explicitly in the previous debate. By indexing beliefs as "S believes that p at t," he makes the epistemic evaluation explicitly time-sensitive: justification is assessed at a particular moment. According to Goldman, this matters because cognitive capacities, background information,

environmental conditions, and defeaters can change over time—and thus the same subject might be justified at one time and not at another, even with the same propositional content (ibid.).

Consider the following example. Let p be the proposition "The wall is green" (Dretske 1970, p. 1015). At time t_1, S walks into a room in normal daylight, with normal vision and no unusual background information. He looks at the wall and forms the belief that p. In these circumstances, visual perception is (for S in that environment) a reliable belief-forming process, so S's belief that p at t_1 is justified. At a later time t_2, however, S acquires new information: a friend tells him, truthfully, that the room is lit with a green-tinted theatrical filter that makes surfaces look green even when they are not. This provides a defeater for p, since it undermines the trustworthiness of the very process—color perception in that lighting—that produced the belief. Even if S continues to believe the same proposition p (that the wall is green), at t_2 the belief is no longer justified, because his updated epistemic situation now includes a strong reason to think that his color perception in this setting is unreliable. The index "at t" thus marks that justificatory status is not a timeless property of p but a time-bound property of S's believing p, sensitive to changes in reliability-relevant factors (A. Goldman and Beddor 2021).

That etiology in Goldman's definition is characterized in terms of a "cognitive belief-forming process" (ibid.)—whereas the term "process" denotes a repeatable method or mechanism that maps certain inputs—perceptual stimulation, memory traces, other beliefs, testimony—onto an output in the form of a belief. Goldman also adds the parenthetical "set of processes" to capture the fact that many beliefs are not produced by a single, simple mechanism (ibid.). Often a belief is the end product of a series of different processes: perception may feed into memory retrieval, which then feeds into an inference, which then yields a belief. The justificatory status of the resulting belief can therefore depend on the reliability of the overall sequence, not merely on one stage in isolation.

On this picture, the belief at issue is thus the doxastic output of such a process (or process combination), and it is the type of process that carries the property of reliability according to Goldman. A particular episode of belief-formation—S's looking at the wall and forming the belief that p at t—is what is known as a *process token* (ibid.); but reliability, understood as a tendency or propensity to generate mostly true outputs over many occasions, is naturally a property of process types. Ordinary examples include types such as normal visual perception in daylight, careful deductive inference, or competent uptake of expert testimony: a given belief-token is justified to the extent that it is produced by a token process that instantiates one or more such truth-conductive types. This move from tokens to types is what allows reliabilism to explain how a belief's etiology can underwrite its justification.

An important consequence of this conception is that a subject can have a justified belief

without knowing they are justified—without any reflective access to the process's reliability. This independence of justification from the subject's perspective has provoked two main lines of sustained criticism. First, the *generality problem* (Conee and Feldman 1998) challenges reliabilism to specify, non-arbitrarily, at which level of description a process should be typed for epistemic evaluation. Any token episode can be characterized as an instance of many different types, each potentially differing in reliability. When Smith sees a maple tree and forms the belief "that is a maple tree," is the relevant process-type "visual perception," "perception under favorable lighting," or "perception of maples on sunny afternoons"? Each specification may yield a different truth-ratio. Since justification depends on the reliability of the *type*, and since typing determines whether the belief counts as justified, the theory requires a principled criterion for selecting the epistemically relevant level of generality (ibid.).[8]

Second, BonJour's *clairvoyance* objection (BonJour 1980) presses whether reliability alone is *sufficient* for justification when the subject lacks evidence or awareness that the process is reliable. Consider Norman, who possesses a perfectly reliable clairvoyant faculty but has no reason to believe he has such a power. One day, spontaneously and without phenomenological cue, Norman forms the belief that the American president is in New York—produced by his reliable clairvoyant faculty. Process reliabilism seems committed to counting Norman's belief as justified, since it results from a highly truth-conductive process. Yet intuitively, Norman is *not* justified: from his perspective, the belief appears arbitrary, and he has no rational basis for trusting it (ibid.). The case suggests that reliability cannot by itself confer justification if the subject's total evidence fails to support the belief.[9]

Summary The externalist frameworks covered ground justification not in a subject's internal perspective but in the objective reliability of belief-forming processes. Whereas internalism makes justification supervene on what is reflectively accessible, externalism evaluates beliefs by their causal etiology: a belief is justified when produced by mechanisms that statistically tend to yield truth. Goldman's process reliabilism systematizes this through temporal indexing ("S believes that p at t"), the token/type distinction, and a historicist approach tying justification to actual belief genealogy. Yet this shift generates persistent difficulties. The generality problem reveals that any token episode can be typed at multiple levels, each with different truth-ratios, yet the theory lacks a principled criterion for selecting the relevant type. The clairvoyance objection presses whether reliability suffices without internal awareness or evidential support.

[8] Goldman has proposed that the right type is fixed by psychologically natural categories (A. Goldman and Beddor 2021), but critics argue this either smuggles in internalist constraints or fails to solve the individuation problem.

[9] Reliabilists have responded by distinguishing *prima facie* from *ultima facie* justification, by restricting the theory to "primal systems" shaped by learning (Lyons 2009), or by incorporating a "no-defeater" clause. Each response, however, reintroduces internalist elements.

The question remains whether externalism can preserve what makes justification distinctively *epistemic* while redefining it as world-connection rather than rational coherence—a tension especially acute, as we shall see, for nonhuman entities, where we must ask whether, e.g., learned connectivity constitutes a truth-conductive process and whether such processes could ground epistemic justification at all.

3.2. Tacit Knowledge

In the previous chapter, we determined what it means for an entity to know something within the context of a propositional statement "S knows that p." But is that sufficient to ground the question "does CLERE know"? Suppose we instead claimed: "CLERE knows how to drive." Does that sentence refer to the same kind of knowledge as "CLERE knows that the capital of Switzerland is Bern"? Proponents of the epistemic theory of tacit knowledge, as we shall see in the following chapter, would deny this: they claim that traditional epistemology has focused too narrowly on explicit, propositional knowledge and thereby overlooked crucial aspects of our knowledge that are non-propositional—things we know that cannot be boiled down to factual statements (riding a bike, speaking a language, recognizing a face).

Whereas the debate around propositional knowledge is guided by the Wittgensteinian notion that "the world is everything that is the case" (Wittgenstein 1922, prop. 1) (the first of his seven propositions in the *Tractatus*) and thus "whereof one cannot speak, thereof one must be silent" (ibid., prop. 7) (the last of his propositions), the epistemology of tacit knowledge begins from the opposite premise. As Polanyi famously puts it: "I shall reconsider human knowledge by starting from the fact that we can know more than we can tell" (Polanyi 1967, p. 4).

Knowing How vs. Knowing That The most notable influence on the development of the theory of tacit knowledge was Gilbert Ryle's distinction between "knowing that" (propositional knowledge) and "knowing how" (practical skill) (Ryle 1946).[10] Ryle argued that possessing factual knowledge is not the same as having the skill or competence to act upon it. One might know all there is to know about how to drive a car—where the accelerator is, where the brake is, how to shift, how to steer—but still not be able to drive it. Or, conversely, one might be an experienced driver capable of difficult maneuvers but still not be able to verbalize how it is that one is able to shift and brake at the exact right time within fractions of a second. This observation led Ryle to challenge what he calls the "intellectualist legend" (Ryle 2009, p. 18): the view that intelligent action is produced by prior inner contemplation of propositions and that knowing how is just a kind of knowing that.

> Philosophers have not done justice to the distinction which is quite familiar to all of us between knowing that something is the case and knowing how to do things. In their theories of knowledge they concentrate on the discovery of truths or facts, and they either ignore the discovery of ways and methods of doing things or else they try to reduce it to the discovery of facts. They assume that intelligence equates

[10] A notion we already find in Aristotle's distinction between *technē* (craft or skill) and *phronēsis* (practical wisdom) on the one hand, and *epistēmē* (theoretical knowledge) on the other; see Aristotle, *Nicomachean Ethics* VI.3–5 (Aristotle 2002).

> with the contemplation of propositions and is exhausted in this contemplation (Ryle 1946, p. 4).

One of the main arguments Ryle puts forward against the intellectualist notion of propositional knowledge is the problem of the *vicious regress* (ibid., p. 2): if an intelligent act required a preceding act of consulting a proposition—as the rule-based approach to knowledge would have it—then, so Ryle argues, that consulting would itself need guidance by another act, generating an infinite regress. To put it using the example from above: if we assume that driving in a difficult situation at time t is an intelligent act that requires the driver to consult a rule or proposition in his mind at $t-1$, then either (1) that act of consultation at $t-1$ is itself an intelligent act, which—by the intellectualist's own argument—must be guided by yet another prior consultation at $t-2$, and so on ad infinitum; or (2) the consultation at $t-1$ is not intelligent but merely mechanical, in which case intelligent performance at t evidently does not require antecedent rule-consultation after all, undercutting the former claim (ibid.).[11]

Therefore, intelligence, so Ryle concludes, is not an inner consultation of truth-bearing propositions but rather the publicly manifest abilities and sensitivities exercised in practice: "[...] Intelligently to do something (whether internally or externally) is not to do two things, one 'in our heads' and the other perhaps in the outside world; it is to do one thing in a certain manner" (Ryle 1946, p. 3). In this spirit, rules and formulas—the "[...] derivative product of theorising [...]" (ibid., p. 12)—are, for Ryle, pedagogical scaffolds, "banisters for toddlers" (ibid., p. 12), useful for training and self-correction but not inner quasi-premises that antecede action.

3.2.1. Polanyi's Theory of Tacit Knowing

Key terms: knowledge vs. knowing, personal knowledge, from-to structure, subsidiary awareness, focal awareness, integration, Gestalt, indwelling, tool use, personal coefficient, rule-following, explicability, inarticulability.

Following Ryle's insight that knowing how is not reducible to knowing that, Michael Polanyi introduced the concept of *tacit knowing* in *Personal Knowledge* (1958) and later elaborated it in *The Tacit Dimension* (1966). Polanyi, originally a chemist, observed that, in practice, scientific discovery and skilled practice rely less on explicit propositions and more on intuition, experiential know-how, and context—aspects of knowledge that resist full articulation in a

[11] A standard attempt to stop Ryle's regress is to posit an inner "rule-applier" (an interpreter or "homunculus") that applies the relevant proposition to the case; but this merely shifts the problem, since the interpreter's application would itself have to be intelligent and so would invite the same regress at the next level (Tanney 2012, pp. 249–276).

propositional sense, hence the famous expression "we can know more than we can tell" (Polanyi 1967, p. 4). The example he often uses to illustrate this notion is the fact that, even though we are able to instantly recognize a familiar face among thousands, we cannot explain how we are able to do that (ibid., pp. 4–6). Such forms of knowing, so Polanyi argues, are carried out tacitly, not in the form of facts or rules one could enumerate in words. Even when we do try to communicate these forms of knowledge (e.g., by describing the face of the perpetrator to a police sketch artist), the very act of matching and selecting features relies on unspecifiable skills—a knowledge that we cannot tell (ibid., pp. 4–6). It is these everyday observations of unspeakable know-how that led Polanyi to the guiding idea that traditional epistemology had overlooked a fundamental epistemic dimension: the realm of personal, experiential understanding that fundamentally underlies our ability to make sense of the world (Polanyi 1998).

From-To Structure At the core of Polanyi's framework is the idea that all acts of knowing have a characteristic *from-to* or *subsidiary-focal* structure (Polanyi 1967, p. 10). We attend *from* a multitude of unstated particulars *to* a coherent whole. In Polanyi's terms, the details we rely on but cannot fully specify (such as the minute features of a face, or the many subtle movements involved in driving a car) constitute our *subsidiary awareness*, while the overall object or pattern they form (the recognizable face, the act of driving itself) is the focus of our attention—our *focal awareness* (ibid., p. 10). According to Polanyi, these two modes of awareness are, in an important sense, incompatible: to the extent that particulars are integrated subsidiarily in terms of something else, they cannot at the same time be grasped in themselves (ibid., pp. 10–11). In other words, when we integrate clues into a whole, we lose sight of the clues as isolated objects. If we shift attention backward—away from the whole and onto the particulars supporting it—the unified pattern collapses.

Polanyi illustrates this with a simple experiment (ibid., p. 11): if one repeats a word over and over and starts focusing on the sounds and movements involved in saying it, the meaning of the word evaporates. What were formerly significant elements (letters or sounds that have meaning when seen as part of a word or sentence) become meaningless fragments when attended to in themselves. Likewise, so Polanyi goes on to illustrate (ibid., p. 12), one will notice that when trying to walk across a tightrope, the only way to reach the other side is by focusing on the goal on the other side of the rope. If one focuses on one's feet instead, the act of balancing is disturbed, and one will most likely fall down. These examples demonstrate Polanyi's key point: tacit knowing is an integrative process. We come to know a comprehensive entity (a pattern, skill, or concept) by integrating many bits of sensory input, cues, and context into a meaningful whole (ibid., pp. 12–13). The knowing is tacit because, while we rely on our awareness of these particulars, we cannot focus on them without breaking the integration. What "we know but cannot tell" are precisely those subsidiarily known contributions that give rise to the overall

meaning.

An important part of this notion can be traced back to the Gestalt theoretic finding that we often perceive forms or configurations without conscious access to the individual elements that compose them.[12] However, Polanyi's account of the tacit goes beyond the Gestalt-psychological claim.

> [...] I am looking at Gestalt, on the contrary, as the outcome of an active shaping of experience performed in the pursuit of knowledge. This shaping or integrating I hold to be the great and indispensable tacit power by which all knowledge is discovered and, once discovered, is held to be true (Polanyi 1967, p. 6).

By emphasizing the active, skillful character of the integration, Polanyi recasts the structure of the Gestalt argument into a logic of tacit thought. He argues that perception and cognition are not passive occurrences but achievements of the knower: an active shaping or integrating of experience performed in the "pursuit of knowledge" (ibid., p. 13). It is important to note in this context that the tacit integrative power that Polanyi describes here is not limited to simple perceptions like recognizing faces or melodies; it operates in all domains of knowing. In fact, Polanyi argues that the highest forms of discovery—the scientist's breakthrough or the artist's creative insight—exhibit the same from-to structure as mundane acts of perception (Polanyi 1998, p. 66). That is why, so he argues, a skilled medical diagnostician can enter a hospital room and sense what is wrong with a patient by picking up on countless subtle cues (such as posture, tone of voice, complexion) without necessarily being able to list each cue at that moment. In practice, the doctor first "knows" the diagnosis tacitly by integrating those clues, and only later might he articulate explicit indicators. In Polanyi's view, such expert knowing is just a more refined version of the same tacit process by which we routinely "rely on our awareness" of particulars for attending to their joint meaning (Polanyi 1967, pp. 10–13).

Indwelling One of Polanyi's central terms for this immersive, absorbed involvement in the particulars of a task or problem is *indwelling* (ibid., pp. 15–18). In tacit knowing, we interiorize parts of the world, assimilating them into our agency and awareness as extensions of ourselves. We dwell in the tools we use, the frameworks we adopt, and even our own bodily capacities in order to attend to something beyond them. "In this sense we can say that when we make a thing function as the proximal term of tacit knowing, we incorporate it in our body—or extend our body to include it—so that we come to dwell in it" (ibid., p. 16). A canonical illustration is Polanyi's discussion of a blind man using a cane to probe his environment: with his skillful

[12] For an overview of Gestalt psychology in visual perception—especially the claim that we often perceive organized wholes (Gestalten) without explicit access to their constituent elements—see Wagemans et al. (2012).

application, attention shifts from the sensations in the hand to the world "beyond the tip" of the cane, which makes the cane not just an external object (a tool) but a medium that connects the blind man to the outside world. In general, Polanyi argues, all tool use (and indeed much of our ordinary bodily engagement) exhibits this pattern of indwelling: "We pour ourselves out into them [sic. the tools] and assimilate them as parts of our own existence. We accept them existentially by dwelling in them" (Polanyi 1998, p. 61).[13]

Through indwelling, meaning is thus carried away from ourselves and established in what we are observing or doing (Polanyi 1967, pp. 13, 15–18). This is why, when we drag the usually tacit elements back into focus (as when we suddenly pay attention to each footstep on the tightrope), we not only disrupt our performance but also lose the meaningful relation to the goal (i.e., fall down). Indwelling expresses the personal participation of the knower in the known: not as a detached representation of propositions, but as a skillful interiorization of clues and tools, guided by the knower's intentions and commitments.

Because of this from-to, interiorizing structure, Polanyi emphasizes that knowing is fundamentally *personal*, *situated*, and *procedural* (Polanyi 1998).[14] All knowing, in his view, requires an agent who contributes commitments, trusts in intuition, and dwells in a framework that cannot be fully objectified (Polanyi 1998). It is from this perspective that his rejection of the ideal of complete objectivity—the "view from nowhere"—must be understood: according to Polanyi, we cannot eliminate the personal coefficients of knowledge without eliminating knowledge itself.

> Any attempt to eliminate this personal coefficient, by laying down precise rules for making or testing assertions of fact, is condemned to futility from the start. For we can derive rules of observation and verification only from examples of factual statements that we have accepted as true before we knew these rules; and in the end the application of our rules will necessarily fall back once more on factual observations, the acceptance of which is an act of personal judgment, unguided by any explicit rules (ibid., p. 267).

Even the most explicit scientific formula or rule, so Polanyi argues, is the product—the "stabilized residual"—of prior skilled integration. They can guide practice only insofar as they are themselves integrated into the tacit process of knowing: "Rules of art can be useful, but they

[13] Tool use is a central theme in philosophy of technology because, in skilled action, tools can become "transparent" and thereby reorganize perception and agency. Phenomenological and post-phenomenological approaches analyze this via embodiment and technological mediation, while other strands connect tools to cognition and social practice (extended mind; STS/actor-network theory). For an accessible survey and further reading, see (Dusek 2006, chs. 5, 12).

[14] This is the reason that he speaks not of "knowledge," the static product, but of "knowing," the ongoing act of integrating and sense-making that combines both forms of know-how and know-that: "I shall always speak of 'knowing', therefore, to cover both practical and theoretical knowledge" (Polanyi 1967, p. 7).

do not determine the practice of an art; they are maxims, which can serve as a guide to an art only if they can be integrated into the practical knowledge of the art. They cannot replace this knowledge." (ibid., p. 52). The core thought that Polanyi expresses here is therefore not merely that rules may be incomplete, but that knowing how to follow a rule is itself a skill. A rule does not apply itself; it must be read as relevant, interpreted in context, timed, and executed under variable conditions—and these enabling competencies typically remain, at least in part, tacit. This is why, so Polanyi argues, it is possible that a skill or practice can get lost if it is not used for some time, even when there is exact written documentation. Take, for example, the recent case of the U.S. National Nuclear Security Administration's (NNSA) attempt to restart production of a certain nuclear warhead (codenamed "FOGBANK") (U.S. Government Accountability Office 2009) that has been dormant for a quarter of a century. Reports describe that, even though there is documentation describing the manufacturing process, "[...] NNSA had lost knowledge of how to manufacture the material because" among other reasons, "[...] almost all staff with expertise on production had retired or left the agency" (ibid., p. 15). Examples like this show, so Polanyi would argue, that when the living practice that sustains these subsidiary skills is interrupted, the explicit record does not by itself preserve the know-how needed to make the rule operative again.

Explicability The last element of Polanyi's account of tacit knowing that is relevant to our understanding is the crucial question of whether the tacit dimension of knowing can, in principle, be made explicit or not. We have seen that Polanyi describes a skillful integration as "[...] achieved by the observance of a set of rules which are not known as such to the person following them" (Polanyi 1998, p. 51). On his view, this does not imply that no rules exist, but rather that it is possible to follow them without consciously knowing the rules as such. The main notion he leaves us with is thus that the tacit component of a skillful integration consists in integrating unspecifiable particulars—the countless tiny cues and adjustments we pick up and make, without being able to list or verbalize them—into a coherent focal whole. But this does not by itself answer the question of in what sense they are unspecifiable. In general, the secondary literature distinguishes between two main ways we might read Polanyi here (Collins 2013, p. 85): a *weak* reading and a *strong* reading. On the weak reading, the tacit dimension is unspecifiable mainly for practical purposes: it is not necessary (and often counterproductive) to articulate the hidden rules in order to use them. On this view, tacit knowledge could be made explicit in principle, at least partially, though doing so may be infeasible or would undermine its smooth exercise. On the strong reading, tacit knowledge is inherently inarticulable: it cannot, by its nature, be captured in propositional form.

Summary Polanyi's theory of tacit knowing begins from the epistemological claim that an exclusive focus on explicit, propositional content overlooks a constitutive dimension of cognition: we routinely achieve understanding and competent performance without being able to state the grounds, steps, or rules by which we do so. His guiding thesis that we "know more than we can tell" is thus not a mere psychological observation but a claim about the structure of knowing.

In order to make this structural claim more explicit, we can thus distinguish four complementary aspects of tacit knowing: a *functional*, *phenomenal*, *semantic*, and *ontological* aspect (Polanyi 1967, pp. 10–13). *Functionally*, tacit knowing has the form of a dependence relation between two "terms": we can know the proximal (subsidiary) term only by relying on it *for attending* to a distal (focal) term (ibid., p. 10). *Phenomenally*, this reliance is reflected in experience: the proximal term is not given as an object of focal attention but is present "in the appearance" of that from which we attend to the distal term (ibid., p. 11). *Semantically*, the relation is one of meaning: subsidiaries are apprehended *in terms of* the joint meaning they contribute to, such that meaning is "displaced" away from ourselves and toward what we are attending to (ibid., pp. 12–13). Finally, *ontologically*, tacit knowing is knowledge of a *comprehensive entity*—a coherent whole constituted by integrating particulars into their joint meaning (for example, the recognized face, the diagnosed condition, or the object reached through the use of a probe) (ibid., p. 13).

The central mechanism underwriting these aspects is Polanyi's *from-to* (subsidiary-focal) structure: in every act of understanding we attend *from* a manifold of subsidiarily relied-upon particulars *to* a coherent focal whole. This explains both why tacit knowledge is hard to articulate and why shifting attention back onto the subsidiaries can destroy the very integration that yields meaning. Tacit knowing is therefore an *active, skillful achievement* rather than a passive reception of facts. With *indwelling*, Polanyi captures how such achievement depends on interiorizing subsidiaries—including tools, bodily capacities, and interpretive frameworks—so that we can attend *through* them to the world (Polanyi 1998).

From this perspective, knowing is fundamentally *personal*, *situated*, and *procedural*: it involves commitment and judgment, and it cannot be purified into a "view from nowhere" without losing what makes it knowledge at all (ibid.). Explicit rules and formulas remain indispensable, yet they function only as guides that must be integrated into practical mastery; following a rule is itself a skill (ibid., p. 52). Finally, Polanyi leaves open to what extent the tacit can be rendered explicit, but his central idea remains: epistemology must account not only for what can be stated, but for the tacit processes by which understanding is achieved.

3.2.2. Collins's Taxonomy of Tacit Knowledge

Key terms: explicability, strings, meaning, language, transmission, relational tacit knowledge, somatic tacit knowledge, collective tacit knowledge, socialization, social Cartesianism, embodiment, shared forms of life.

According to Collins, a pervasive—often unexamined—assumption shaped the debate on tacit knowledge in the period when Ryle and Polanyi were developing their accounts: namely, that tacit knowledge is the difficult, obscure notion, whereas explicit (propositional) knowledge is the easy, straightforward one.

> Though the tension between tacit and explicit goes back at least as far as the Greeks, it was modernism in general and the computer revolution in particular that made the explicit seem easy and the tacit seem obscure. But nearly the entire history of the universe, and that includes the parts played by animals and the first humans, consists of things going along quite nicely without anyone telling anything to anything or anyone [...]. There is, then, nothing strange about things being done but not being told—it is normal life. What is strange is that anything can be told (Collins 2013, p. 7).

Collins contends that when we frame matters with the slogan that "all knowledge is either tacit or rooted in tacit knowledge," the explicit appears dependent on the tacit; yet with respect to our *idea* of the tacit, the dependence is reversed: the very idea of the tacit is "parasitic on the idea of the explicit" (ibid., p. 7). Only once explicability is taken as the ordinary condition does tacitness come to seem hard to understand and transmit. Hence, twentieth-century epistemologists needed the category of tacit knowledge precisely because propositional knowledge had become culturally privileged and misleadingly thought of as the "easy" form of knowing. "[...] Polanyi was writing in the postwar period when the world was a very different place. In particular, it was a world dominated by science and the belief that logic, observation, and experiment were, without question, the preeminent ways of obtaining sure knowledge" (ibid., p. 148).

With this misconception in mind, Collins tightens the, so far, only loosely defined distinction between implicit and explicit knowledge by introducing the term "string" to denote the physical material that propositional knowledge is dependent on—i.e., transmitted on. On his account, a string is a physically realized pattern (marks on paper, sound waves, magnetic states, bits in memory, brushstrokes) that can be transmitted and interpreted without containing any meaning by itself.[15] He thus sharply distinguishes strings from language: the former is "just a physical

[15]The term "string" as Collins uses it here is thus not to be confused with "string" as a computer

object," whereas the latter is "a set of meanings located in a society" (Collins 2013, p. 9). The example he uses to illustrate this is da Vinci's Mona Lisa: Without the onlooker, without someone ascribing meaning to the brushstrokes on the canvas, there is no smile: "Examine the painting very carefully and you will find that the smile consists of nothing but paint. But the paint is not the smile. The smile is some combination of the paint and the person looking" (ibid., p. 45).

According to Collins, propositional knowledge is thus the part of our knowledge that can "ride on strings" and be transmitted through middle persons or middle things (books, datasets, code, recordings), while the tacit is that which cannot be or is not transmitted with strings and needs direct contact (ibid., 15ff.). With respect to our question from above, i.e., in what sense the tacit is unspecifiable, he argues that the term "tacit knowledge" actually encompasses three different phenomena that correspond to this question: "[...] there are three major reasons why tacit knowledge is not explicated; therefore, there are three major types of tacit knowledge" (ibid., p. 1): *relational*, *somatic*, and *collective* tacit knowledge.

3.2.2.1. Relational Tacit Knowledge

The first type, relational tacit knowledge (RTK) (Collins 2013, pp. 85–98), corresponds to the weak reading of tacitness. It is a form of knowing that is only tacit because of the social relations from which it emerges. Within this context, Collins distinguishes four types of RTK: First, there is *concealed knowledge* (ibid., p. 91) that could, in principle, be made explicit but is not because someone or something—a person, an institution, a social convention—deliberately keeps it hidden. In the case of apprenticeship, the knowledge-forming process between master and apprentice would therefore be tacit insofar as some "tricks of the trade," as Collins puts it, are deliberately kept back[16] because "[...] humiliation of the uninitiated seems part of the ritual and the power relation" (Collins 2013, p. 92). Second, even if no one is deliberately obscuring the process, there is *mismatched saliences* (ibid., p. 95): A wants to teach B, but what is important to A is not salient to B (or A wrongly assumes background that B does not have). Third, there is *ostensive knowledge* (ibid., p. 93), where the only effective way to transmit the knowledge is to point, demonstrate, or otherwise show an object or practice, because a purely verbal description would be too unwieldy or would fail to secure the right

scientist or a physicist would understand it. In computer science, a string is typically a finite ordered sequence of symbols drawn from an alphabet (e.g., characters) (Hopcroft, Motwani, and Ullman 2006, p. 1); in physics, "string theory" uses the metaphor of one-dimensional extended objects ("strings") as the fundamental constituents (Zwiebach 2009, p. 1). Collins's notion is broader than both: it is, as he notes, akin to the computing usage, but "[...] more general still, including [...] the physical medium on which the information contained in the pattern is expressed" (Collins 2013, p. 10).

[16] In some cultures, learning a skill within an apprenticeship is described as "stealing the master's secrets." See (Collins 2013, p. 93).

uptake. Fourth, there is *unrecognized knowledge* (ibid., p. 95), where A reliably does something "the right way" without realizing which detail does the causal work and thus cannot tell B what to copy; only later is the operative parameter identified and codified (e.g., a workshop "rule of thumb").

To summarize, RTK describes a form of knowledge that could, in principle, be made explicit but is not "[...] because of the contingencies of human relationships, history, tradition, and logistics" (ibid., p. 98). It is, as Collins puts it, a "[...] ragbag category—but a vital one" (ibid., p. 92), since there will always be knowledge that is not made explicit for these contingent reasons, and it therefore "[...] will be an ever-present feature of the domain of knowledge as it is encountered even though its content is continually changing" (ibid., p. 98).

3.2.2.2. Somatic Tacit Knowledge

Second, Collins argues that tacit knowledge often eludes explicit formulation for reasons rooted in the organization of the human body and nervous system; a category he denotes as *somatic tacit knowledge* (STK) (Collins 2013, pp. 99–117). In contrast to the positions within the debate of *embodiment*[17] that see the human body as the key to understanding the tacit dimension of knowing, Collins advocates demystifying this fixation on the human body as something special.

> In sum, there is nothing philosophically profound about Somatic tacit knowledge, and its appearance of mystery is present only because of the tension of the tacit with the explicit: if we did not feel pulled toward trying to say what we do, and if we did not make the mistake of thinking this is central to the understanding of knowledge, we would find nothing strange about our brains' and bodies' abilities to do the things we call tacit. And that is why too much concentration on the body as the seat of the tacit takes one away from a proper understanding of the idea (Collins 2013, p. 117).

Collins argues that the fixation of the debate on the body is to be understood as a counterweight reaction to the technological advances from the early symbolic AI movement (see Section 2.1). The "bulwark of bodily competence," he argues, was erected to resist over-logicization, an idea that must now be reconsidered (ibid., p. 117).

The second, semantic dimension of tacit knowledge thus does not correspond to a strong reading of the tacit but acts as a kind of middle position in Collins's typology. He distinguishes

[17] In the embodiment debate, cognition and skill are treated as grounded in a subject's sensorimotor capacities and practical engagement with an environment rather than as purely inner rule-following. In the context of tacit knowledge, this motivates the idea that competent performance often depends on bodily know-how that is difficult (and sometimes pointless) to render as explicit instructions. For an accessible overview of the main positions and debates, see Shapiro (2019).

between two main subtypes: *somatic-limit tacit knowledge* (SLTK) (Collins 2013, p. 101) and *somatic-affordance tacit knowledge* (SATK) (ibid., p. 109). SLTK names cases where making the knowledge explicit would be of no practical use to humans because of intrinsic physiological or neurological limitations (no explicit recipe could be followed effectively), and SATK covers the converse cases where articulation of the knowledge is unnecessary due to the affordances of the human body or mind (no explicit recipe is necessary). To illustrate the former, Collins uses Polanyi's paradigmatic example of riding a bike (ibid., p. 101): he argues that the reason why cases like this constitute a form of tacit knowledge is not because there is something about activities like riding a bike that is in principle incompatible with explicit rules but because explicit rules simply cannot be consulted and executed in sufficient time. If we were to ride a bike on a small asteroid with almost zero gravity, where everything happens much slower, "[...] balancing on a bike would be like assembling flat-pack furniture: as we began to fall to the left or the right we would consult a booklet and slowly adjust the angle of steering according to the instructions for remaining upright" (ibid., p. 100). And the same goes for SATK. Imagine, as Collins does, a manufacturing company with an old warehouseman who knows where everything is (ibid., p. 94): if we were to ask him to draw an exact map of the warehouse containing the location of every item, he would not be able to do it; but, if we were to ask him for a specific part, perhaps describing its function or its shape and size, "[...] something in his brain and, perhaps body, will bring him to the place where it is kept" (ibid., p. 94). The warehouseman, so Collins argues, has somatic tacit knowledge of the warehouse, and the reason why he is not able to explicate this knowledge is because he does not need to in order to function in his position. His brain is simply structured in a different way than, let us say, a computer in his place.

In summary, if what we mean by "explicate" is that something can be expressed in scientific terms of causal sequences—and, therefore, as we shall see (Section 4.2.2), be (in principle) reproduced by machines—then STK does not constitute a tacit form of knowledge in the strong sense. The limits to reproducing STK are not limits of principle but limits of practical ability and the affordance of different materials: trying to reproduce human capacities with the wrong stuff, so Collins argues, is like "[...] making a slide rule out of elastic bands or a rocket out of ice [...]" (ibid., p. 119). The endeavor will fail not because the capacities are in principle beyond explication or mechanization, but because the chosen substrates do not afford the requisite causal organization.

3.2.2.3. Collective Tacit Knowledge

Third, Collins identifies a form of tacitness that is neither a mere artifact of the contingencies of human relations, history, and traditions (RTK), nor grounded primarily in the limits and affordances of human bodies and minds (STK). He calls this *collective tacit knowledge*

(CTK) (Collins 2013, pp. 119–138). Here, the factor that earlier examples (e.g., riding a bike; apprenticeship) bracketed comes to the fore: the collective dimension of knowledge. From the perspective of CTK, riding a bike is not only a matter of balancing but of negotiating traffic, which—so Collins argues—calls upon a different kind of know-how, one that depends on the social context: "[...] it involves knowing how to make eye contact with drivers at busy junctions in just the way necessary to assure a safe passage and not to invite an unwanted response. And it involves understanding how differently these conventions will be executed in different locations" (ibid., p. 121). Thus, riding a bike in Amsterdam requires a different grasp of unspoken, collective norms than riding a bike in Delhi.

According to Collins, navigating traffic is thus a polymorphic practice in which correct performance and interpretation depend on context and on others' shifting expectations; any fixed set of rules—any full explicitation—would at best be a frozen snapshot of past practice. Take Searle's Chinese Room (see Chapter 3) as an example: Collins argues that even if we assumed a perfect lookup table that contained a complete dictionary from one language to another—setting aside the practical impossibility of that assumption[18]—that dictionary would still be static in the crucial sense that it could not track the ongoing changes of a living language. There is, unlike in STK, no "low-gravity" method for speaking Chinese: even a complete dictionary cannot substitute the conditions under which humans acquire and sustain fluency—namely, socialization, i.e., the internalization of CTK (Collins 2013, p. 127). Within Collins's work, this is referred to as the *socialization problem*: we can describe the circumstances under which linguistic fluency is acquired but lack a specification of the mechanism by which individuals remain "in touch" with society and its continual change. It is this problem that leads Collins to adopt what he calls *social Cartesianism* (ibid., p. 125)—the methodological claim that the collectivity, not the individual, is the primary locus of knowledge—and to ground his epistemic notion of CTK in that stance.

Collins argues that there has not only been a cultural bias toward the propositional side of knowledge, but also—especially in Western academic communities—toward individualism (ibid., p. 131). The notion of the individual mind as the basic unit of analysis is so deeply entrenched in Western academia that it is hard to imagine the collectivity as the location of anything—even though, so Collins argues, there is nothing special about this notion, simply because there is nothing special about the boundaries of the individual's brain.

[18] To illustrate the logistical point, follow Ned Block's suggestion to simplify the thought experiment by switching to an *English Room* (Block 1981). Assume (1) a finite space of possible questions and (2) that any question must be at most 100 characters long; allow answers up to the same length, yielding 200 characters per question-answer pair. With a generous "typewriter" alphabet of about 100 symbols (letters, digits, punctuation, space), the space of 200-character strings is $100^{200} = 10^{400}$. Compare this with the estimated number of atoms in the observable universe ($\sim 10^{80}$) to appreciate by what astronomically large factor such a lookup table would exceed any conceivable storage capacity—rendering it practically impossible.

> My brain's connections do not stop at the boundary of my head, because they are not limited by the connections found within the grey matter; my brain's neurons are connected to the neurons of every other brain with which it is "in touch" via my five senses. There is nothing even remotely strange about saying that the seat of knowledge is the collectivity of brains because the collectivity of brains is just as much a "thing" as my individual brain is a "thing"; my brain is a collection of neurons separated by (if we were to examine them on an atomic scale) huge distances so the distance between brains in the collectivity is no obstacle to their comprising one "thing" between them (Collins 2013, p. 132).

This notion of the collectivity of brains can be understood, so Collins argues, without any metaphysical overhead: it merely represents the other side to the idea of a single brain: a web of connections whose "weights" shift whenever social and technological life is reorganized.[19] Knowledge, on this view, is located in the collective insofar as brains are linked by speech and practice: "The metaphysically bashful can just think of all brains linked by speech as making up one big neural net" (Collins 2013, p. 132). The individual, in turn, is a temporary, "leaky repository" of collective knowledge whose connection to the CTK—such as language—degrades without sustained contact. "We have, then, a picture of the individual as a parasite on the social group, sucking up social knowledge from the super-organism; stop sucking and the knowledge gradually degrades—that is, its match with the collective's knowledge gradually weakens" (ibid., p. 132).

With regard to the previous notion of STK, the question thus arises: if CTK is located in the collectivity, what role does embodiment play in accessing and shaping it? Collins offers two answers: First, he argues that the content of CTK is (at least partly) a function of a species-typical body—hence Wittgenstein's famous claim that even if a lion could speak, we could not understand him, since its form of life (and thus its conceptual world) would be differently structured than ours (a position Collins calls the *social embodiment thesis* (ibid., p. 135))—and second, although bodily forms help constitute the content of CTK, so Collins argues, participation in the collective can be sustained with only the minimum apparatus for discourse—i.e., much of the CTK is carried by language itself (what he calls the *minimal embodiment thesis* (ibid., p. 136)). The latter would explain, e.g., how someone like Stephen Hawking, communicating only via prosthetic means, could continue to acquire new elements of CTK through "linguistic immersion alone."

In contrast to RTK and STK, Collins's CTK cannot be made explicit so long as the socialization problem remains unsolved. The core difficulty, in brief, is that we lack a mechanistic account of

[19] The analogy to artificial neural networks is strong here. Collins himself argues that "[...] the very concept of the neural net shows us how to think about it this way without invoking anything mysterious like 'collective consciousness' " (Collins 2013, p. 132).

how individuals are inducted into—and kept current with—the living collectivity. Although we can describe the circumstances of acquisition, we cannot specify the procedure or engineer artificial learners that would "parasitize" the collective as humans do. However, does this correspond to a strong reading of tacitness? If we look back on Collins' notion of STK, we find that the fact that only the human body-brain combination sustains CTK is not probative for the epistemological question; to treat it as such would mistake a contingent implementation constraint for a metaphysical claim of human exclusivity.

> The fact that acquiring collective tacit knowledge via practice is sometimes more and sometimes less efficient than acquiring it through discourse alone is, once more, a matter of the nature of humans not the nature of knowledge. No human body is good at acquiring lots of practical skills and some human bodies are better at acquiring some practical skills than other human bodies. But these different efficiencies are not "metaphysically significant," as one might say (ibid., p. 138).

At the epistemological level, CTK is a matter of participation in a form of life and is therefore substrate-independent: Accordingly, we must infer that any system that could be socialized to those standards would count as a bearer of CTK—even if we cannot yet envision its nonhuman implementation.

3.3. Summary

This chapter developed the epistemic vocabulary needed to assess whether a subsymbolic system such as CLERE can plausibly be attributed "knowledge" in more than a merely metaphorical sense. Starting from the distinction between the symbolic (cognitive) and subsymbolic (connectivist) paradigms, we used Searle's *Chinese Room* to articulate a classical worry about rule-based symbol manipulation: that syntactic competence can be behaviorally impressive while remaining semantically empty (Chapter 3). Against this background, the guiding motivation of the chapter was to ask how the epistemological picture changes once competent performance is not implemented via an explicit rulebook but via learned internal structure (associations, representations, connectivity).

The first half of the chapter (Section 3.1) reconstructed the standard template for *propositional* knowledge in the form of the justified-true-belief (JTB) analysis and the historical debate about what justification amounts to. Rationalist and empiricist strategies were introduced as contrasting accounts of epistemic support (a priori structure vs. experience), with Hume's problems of causality and induction sharpening the limits of purely empirical justification. Kant's synthesis then served as a bridge: knowledge begins with experience but depends on a priori forms and categories that make experience objectively tractable. This set the stage for the decisive modern pressure on JTB: Gettier-style cases show that internal adequacy of reasons can still yield true belief by luck, so that JTB is not sufficient for knowledge.

The Gettier problem (Section 3.1.2) motivated the central structural division of contemporary epistemology into *internalist* (Section 3.1.2.1) and *externalist* (Section 3.1.2.2) families. Internalism was presented as a supervenience claim: justification is fixed by the subject's epistemic perspective (illustrated by the evil-demon/matrix scenario). We then distinguished access internalism (reflective availability of reasons) from mentalism (justification fixed by mental states whether or not reflectively accessible). Bayesian epistemology was introduced as a formally regimented internalist framework: degrees of belief (credences) are constrained by probabilism and updated by conditionalization, shifting justification from a static "ledger of reasons" to a norm of coherent evidential updating. Externalism, by contrast, was developed via the demand that justification be *truth-conductive* in the actual world, leading to Goldman's process reliabilism: beliefs are justified (in the relevant sense) when produced by reliable belief-forming processes, evaluated in a time-sensitive way.

The second half of the chapter (Section 3.2) broadened the target beyond "knowing that" to "knowing how." Ryle's distinction between knowing how and knowing that, together with his regress argument against intellectualism, motivated the need for a non-propositional dimension of epistemic assessment. Polanyi's theory of tacit knowing (Section 3.2.1) then

provided a structural account of such competence: the from-to (subsidiary–focal) integration, indwelling, and the personal/situated character of knowing, culminating in the question of explicability (weak vs. strong readings of tacitness). Finally, Collins's (Section 3.2.2) intervention reframed tacit knowledge by emphasizing the material carrier of the explicit ("strings") and by distinguishing three reasons for non-explication—relational, somatic, and collective tacit knowledge—thereby foregrounding the socialization problem and the locus of knowledge in a form of life.

Taken together, the chapter yields a structured set of criteria that can be operationalized in the next chapter's application to CLERE: (i) propositional knowledge assessed under internalist (including Bayesian) and externalist (reliabilist) conditions, and (ii) tacit knowledge assessed via Polanyian integration/indwelling and via Collins's taxonomy, especially the collective dimension that ties competence to ongoing participation in social practice.

Framework / position	**Epistemic criterion / central claim**	**Guiding questions for CLERE (next chapter)**
Propositional knowledge ("knowing that")		
JTB + classic justification debate (Descartes/Hume/ Kant)	Knowledge minimally: true belief + justification; justification contested (a priori structure, experience/habit, Kantian synthesis).	Can the JTB schema be applied to CLERE? If so, what could "belief" and "justification" plausibly amount to for a non-human entity, and which parts of the classical theory must be revised (if any)?
Gettier problem	JTB can be true by luck; internal adequacy of reasons does not guarantee knowledge.	Given that pre-Gettier JTB is incomplete, which post-Gettier strategy is most appropriate for evaluating CLERE's propositional status: an internalist reconstruction of belief/justification, an externalist assessment of reliability (truth-conductiveness), or both as complementary lenses?
Internalism (access; mentalism; Bayesian credences)	Justification fixed by the subject's internal epistemic profile (reasons/mental states); Bayesian norm: coherent credences + rule-governed updating.	Is there an internalist reading of CLERE's outputs that treats them as belief-like states (e.g. graded confidence) and ties them to an internal notion of justification?
Externalism (truth-conductiveness; process reliabilism)	Justification/knowledge requires an appropriate world-connection: belief-forming processes must be reliably truth-conductive across relevant conditions.	If CLERE is approached as a belief-forming process rather than as a reason-giving subject, what would count as sufficient reliability for knowledge attribution? Which features of CLERE's development and use would have to anchor its outputs in the world in a truth-conductive way?
Tacit knowledge ("knowing how")		
Ryle (knowing how $\neq$ knowing that)	Intelligent performance is not (in general) prior rule-consultation; intellectualism risks regress; skill is "doing in the right way."	How does Ryle's knowing how/knowing that distinction reorient the discussion from doxastic states to competence and thereby set the stage for treating CLERE's "tacit" dimensions primarily via the Polanyi framework?
Polanyi (tacit knowing)	From–to (subsidiary–focal) integration; indwelling; personal/situated judgment; weak vs. strong readings of explicability.	Does CLERE's non-explicit organization support more than a merely superficial analogy to Polanyian tacit knowing? How does the analogy look when examined along Polanyi's three axes of (i) from–to integration, (ii) indwelling, and (iii) explicability?
Collins (strings + RTK/STK/CTK)	Explicit rides on physical "strings"; tacit splits into: RTK (contingencies of relations), STK (bodily limits/affordances), CTK (form of life; socialization problem).	Which kinds of non-explicitness are at issue in CLERE's case: relational (disclosure/uptake), somatic (material and temporal constraints on usable explicitness), or collective (dependence on social embedding)? In particular, how (if at all) does Collins's socialization problem bear on CLERE and motivate a shift toward social epistemology?

Table 3.1.: Compact overview of the epistemic frameworks developed in this chapter and the guiding questions they raise for the assessment of CLERE in the next chapter.

4. CLERE's Epistemic Status

In the last two chapters, we prepared the ground for a systematic assessment of CLERE's epistemic standing from two complementary directions. First, we distilled the technical development from symbolic rule systems to subsymbolic, data-driven learners into an explicit model of how CLERE achieves its competence: a system trained on selected (and never simply "given") data, improved by error-driven self-modification, and stabilized in the form of distributed, high-dimensional internal representations (Section 2.10). Second, we reconstructed the epistemological vocabulary needed to evaluate whether such a system can plausibly be said to "know" anything, developing criteria for propositional knowledge (JTB and its post-Gettier revisions under internalism/externalism) (Section 3.1) as well as for tacit knowledge (from Ryle and Polanyi to Collins's taxonomy and the socialization problem) (Section 3.2).

In the following chapter, we bring these strands together in a structured assessment of CLERE's epistemic status: under which sense(s) of "knowing" would CLERE qualify as a knower, what conceptual adjustments would that require, and where do the relevant epistemic predicates plausibly attach—to the system itself, to its training history, or to the socio-technical practice within which it is deployed? We begin with propositional knowledge. Starting from the pre-Gettier JTB schema, we ask whether its central terms—truth, belief, and justification—can be applied to CLERE without a category mistake, and, if not, how they would have to be recast for a non-human system. The Gettier challenge then motivates two complementary strategies. An internalist route (Section 4.1.1) explores how belief- and justification-like notions might be reconstructed in terms of confidence, evidential support, and the system's own training-sensitive organization, while keeping in view the black-box character of ANN processing. An externalist route, by contrast, treats CLERE less as a bearer of reasons and more as a candidate belief-forming process, asking what would count as sufficient truth-conduciveness and which features of development and deployment would have to anchor its outputs reliably in the world.

In the second part, we then turn to tacit knowledge. Here the central question is not whether CLERE can endorse propositions, but whether its non-explicit organization and performance profile support more than a superficial Polanyian analogy. We test this by examining the analogy along the three axes developed in Chapter 3: the from-to structure of integration, the role of indwelling and embodied participation, and the scope and limits of explicability. This analysis is meant to clarify a pervasive temptation in contemporary AI discourse: to slide from opacity or uncodifiability to the claim that a system therefore possesses tacit knowledge in a robust, subject-involving sense. Finally, we draw on Collins's taxonomy to distinguish different sources and locations of non-explicitness (relational, somatic, and collective), and to foreground the socialization problem as the most demanding constraint on attributions of collective tacit knowledge to CLERE.

4.1. Propositional Knowledge

Let us start at the beginning and ask whether CLERE can be said to have knowledge in the pre-Gettier JTB sense by simply applying the definition:

> CLERE knows that p iff: (1) p is true, (2) CLERE believes that p, and (3) this belief is justified.

If we recall the epistemological debate to which this analysis of knowledge is connected, it quickly becomes evident that applying it to CLERE risks a category mistake (anthropomorphic fallacy, Chapter 1). The three conditions of knowledge were formulated against the background of conscious human subjects; they are meant to articulate what it is for a person to be in a cognitive state of knowing. While it might be possible to argue for p to be true regardless of the subject holding the belief,[1] the second and third conditions, as they have been debated in traditional epistemology, clearly were not conceptualized with the notion of a non-human subject of knowing in mind.

In the classical framework, belief is a doxastic attitude of a subject: to believe that p is to take p to be the case, to be disposed to rely on p in reasoning and action, and (on many accounts) to stand in a phenomenally or reflectively accessible state regarding p (Section 3.1). A common distinction within JTB is to read "belief" either in a weak sense, where one might believe something "[...] by virtue of being pretty confident that it's probably true [...]" (Ichikawa and Steup 2024), or in a strong sense ("outright belief") where it is not enough for S to be "pretty sure" or "confident" that p is true but "[...] it is something closer to a commitment or a being sure [...]" (ibid.). Likewise, the justification condition is typically spelled out in terms of the subject's evidence, reasons, or reliable cognitive processes—features that presuppose an epistemic perspective from which one can possess, assess, and (at least in principle) articulate one's grounds for p.

In either case, there is no obvious way in which CLERE can be said to believe or justify that p. CLERE has no doxastic perspective from which p could be endorsed, relied upon, or treated as a commitment; it merely transforms inputs into outputs according to its learned parameters. As we shall see (Section 4.1.1), there are ways in which we can adapt these two conditions in order to arrive at a more flexible notion of JTB. But if we retain the belief and justification

[1] In contemporary debates, the main families of truth theories include: (i) *correspondence theories*, on which a proposition is true if and only if it corresponds to a mind-independent fact or state of affairs; (ii) *coherence theories*, which identify truth with belonging to a maximally coherent set of beliefs; (iii) *pragmatist accounts*, which tie truth to what would be stably accepted under ideal inquiry or to what "works" in practice; and (iv) *deflationary* or *minimalist views*, which treat the truth predicate as a logical device (so that saying "it is true that p" adds no content beyond simply asserting p). For a comprehensive overview, see (Kirkham 1992). We will remain neutral between them here, assuming only that some propositions are in fact true and others false.

conditions in their literal, classical sense, the idea that CLERE "knows" that p cannot be upheld without a substantial revision of what we mean by knowledge.

That being said, the classical notion of JTB, as we have seen, has been shown to be incomplete by Gettier-style counterexamples (Section 3.1.2), and especially the third condition has become the target of extensive revisions and alternative analyses in 20th- and 21st-century epistemology. With regard to the question at hand, the subsequent split into internalism and externalism provides us with two corresponding ways to analyze the notion of CLERE's propositional knowledge: we can either focus on the knowledge-generating *process* within the CLERE system and ask whether this process reliably generates true beliefs, or we can analyze in what sense CLERE can be said to "believe" and in what sense its beliefs are tied to a justification process.

4.1.1. Internalist View

The first problem we run into when analyzing CLERE from an internalist perspective is what is usually referred to as the *black box problem* (Räuker et al. 2022): due to the nature of ANNs, the knowledge-generating process that takes place between input and output—implemented by a vast number of parameters distributed across a very high-dimensional space (see Section 2.10)—is, at a certain level, opaque to a human observer. Although we have a conceptual (global) understanding of the architecture (we know how CLERE's network is set up, how its loss function and optimization procedure are defined, and how training proceeds in general[2]), we lack a local, fine-grained mechanistic understanding of why a given input leads to a given output. In particular, we cannot say, in any illuminating way, why input X_1 results in a change $\Delta w_i = y$ in some parameter w_i, while X_2 leads to a change $\Delta w_j = z$ in w_j.

A critical voice might object here that we also do not have this kind of mechanistic understanding of the human brain. One might even argue that, at the level of synapses and neurons, we have considerably less detailed knowledge of how the human brain works than we have of how CLERE's artificial neural network is structured (Lichtman and Denk 2011): since, in the latter case, we have complete access to the architecture with all its parameters and can therefore perform controlled interventions and experiments, whereas the former is only partially accessible and, for the most part, remains epistemically opaque at the level of its fine-grained neural microstructure (ibid.). The difference between CLERE and a human epistemic subject,

[2]It is important to note that in many real-world deployments this "global" understanding is, as we shall see in Section 4.2.2, itself unavailable, because key aspects of a model (architecture details, training data, parameter values, or even parts of the training procedure) are treated as proprietary information and deliberately withheld as trade secrets (Burrell 2016). For the present epistemological discussion, however, we set this contingent constraint aside and focus instead on the more intrinsic opacity that arises even when the basic design choices are known.

however—and this is the crucial point—is that in the human case, we do not treat the brain as the epistemic subject. In the JTB formula (Definition (JTB)), S does not refer to the brain as a neural network but to the person as a whole—something that transcends its neurobiological foundations.[3]

With respect to the notion of justification, we can thus note that a human can (in principle) report on the reasons why they believe p, what experiences support their belief, and how their belief coheres with their other beliefs, because there is, next to the brain (the neural network), another level that we can address as the epistemic subject (the person). With CLERE, we do not have this option: there is no entity that transcends the network in the same way. If we are to describe CLERE as an epistemic subject of propositional knowledge, we must therefore adapt the terms to fit the reference. We have shown that there are (at least) three main ways to do so within the internalist framework.

(1) If *access internalism* is the target, then justification requires that the relevant J-factors be (at least in principle) reflectively accessible to the subject. On that construal, CLERE seems like a particularly hard case: even if the network contains rich internal structure (activations, latent representations, stored statistical regularities), there is no straightforward analogue of the subject's deliberative standpoint in which these states are available as reasons. Although we can build observer-facing access to internal structure (for example through interpretability methods), such access is primarily the observers', not CLERE's; it does not by itself amount to the system *having* reasons in the accessibilist sense.

(2) *Mentalism*, by contrast, weakens the access requirement: justificatory status supervenes on the subject's mental (or mental-state-like) profile whether or not the subject can explicitly bring that profile to reflective awareness. This makes mentalism the more natural internalist family to test on a system like CLERE. The immediate question then becomes how far we are willing to treat CLERE's internal representational states as playing the role that beliefs, seemings, and memories play in a human mentalist picture. Even if we grant a functional analogue here, a residual worry remains: mentalism still evaluates a subject's internal profile, and it is not obvious that a merely subpersonal cascade of activations can constitute the kind of epistemic standpoint that internalists typically have in view.

(3) The most fruitful way to make this issue precise, however, is to move to the *Bayesian* regimentation of an "internal profile." The Bayesian framework preserves the internalist strategy (justification fixed "from within"), while replacing the opaque talk of "reasons" with a formally tractable structure of graded attitudes and rule-governed updates. Applied to CLERE, this

[3]For the sake of completeness and without going into too much detail, we note that some strongly reductive or neurocentric positions treat persons as nothing over and above their brains, and accordingly shift psychological or epistemic predicates (believing, judging, knowing) from the human being to the brain or its parts. For an overview, see Kim 1996.

suggests the following reading: its "belief" that p is given by the probability it assigns to p conditional on its training and current input data. The conditional probabilities encoded in its output distribution then play the role that credences play for a human Bayesian agent. During training, CLERE's parameters are adjusted so that its output probabilities approximate the empirical frequencies and conditional dependencies in the training corpus. In deployment, these parameters then implement something like a fixed prior over continuations[4] which can be modulated by new evidence.

From an internalist, Bayesian standpoint, CLERE's belief regarding a proposition could thus be reconstrued as its treating p as likely (or unlikely) given its internalized model of the data. Justification, correspondingly, would consist in the fact that these likelihood assignments are the result of a disciplined learning procedure, i.e., stochastic gradient descent on a loss function that encodes a notion of error.

By applying Bayes to CLERE, we thus obtain an internalist reading of its "belief states" in terms of graded credences and rational updating, from which we could, in principle, reconstruct CLERE as a quasi-Bayesian agent whose attitudes toward propositions are encoded in its output probabilities and their training-induced justification. There are already contemporary approaches that *operationalize* this notion as a concrete evaluation tool for classification tasks: Virani, Iyer, and Yang (2019), for instance, propose so-called *epistemic classifiers*: very roughly, a model's output on an input x is treated as its "belief" that x is of class y, while the local neighborhood of x in the input and latent spaces of the network is used to construct a "justification" from the training data. If the nearby training points provide clean and consistent support for the predicted class, the classifier labels its prediction "I know" (IK); if the local evidence is mixed, it labels it "I may know" (IMK); and if there is no appropriate support at all—e.g., in regions of extrapolation—it labels the case "I don't know" (IDK) (ibid.). In Bayesian terms, the model's conditional output probabilities thus still play the role of credences, but these epistemic labels encode how well those credences are backed by accessible evidence in the learned representation space. With the notion of "epistemic classifiers," the strength of a model's putative knowledge claim about x is thus constrained not only by its numerical confidence—as with traditional benchmarks—but also by the quality of the internal support it can muster from its own training history and internal states.

Summary To summarize, from an internalist perspective the key distinction in the comparison between humans and ANNs like CLERE is not primarily at the level of micro-mechanisms, but

[4] In strict Bayesian terms, the prior is specified before observing any data, whereas CLERE's parameters are already an informed point estimate (or implicit posterior) obtained from pre-training data. "Prior over continuations" is therefore heuristic here: it means only that, at deployment time, the fixed parameter configuration functions as a background expectation that is then conditioned on the particular prompt; cf. Wasserman (2004, Sec. 11.2).

at the level of *who* or *what* counts as the epistemic subject. Internalists do not require that a knower understand the neural or computational processes that produce a belief; what matters is that the subject has access to their *reasons* for believing p: experiences, memories, and other beliefs that they can, in principle, cite, assess, and revise. This is a personal-level notion of justification, situated in what Sellars (1956) and others call the "space of reasons," rather than in the subpersonal "space of causes" where neural firings or weight updates occur (ibid., §36).

For CLERE, by contrast, the entire epistemic process is located on a subpersonal level. CLERE maps inputs to outputs via a high-dimensional cascade of activations and weight updates, and even if we could reconstruct this causal chain in detail, there is no further "someone" who takes certain internal states as reasons for others.[5] We can, as we have seen, describe CLERE's outputs in belief-like and justification-like terms; however, the question remains open whether such modified terms still justify knowledge from an internalist perspective or whether justificatory status essentially presupposes an agent that can occupy a standpoint in the space of reasons.

4.1.2. Externalist View

The second perspective we can assume when applying the post-Gettier JTB to CLERE is to raise the question of whether the internal knowledge-generating process—however justifiable that label may be—is tied to external factors in a way that can be said to reliably generate true beliefs. For if that is not the case, so the externalist argument goes, it does not matter whether the internal process is accessible to the subject or transparent; the result still does not warrant the term knowledge. To illustrate this perspective, let us consider the following example.

In a 2016 study, Wu and Zhang (2016) trained a neural network on roughly 1,850 ID photographs labeled either as "criminal" or "non-criminal." From that training, the ANN learned to distinguish between those two labels on the basis of certain low-level facial features such as the distances between eyes or the shape of the mouth. According to the intentions of the authors,

[5] A technically interesting question here is to what extent one could engineer something closer to an "internal profile" by adding explicit self-models—e.g., second-order "meta-networks" that take a first-order network (or its weights) as input and learn compact embeddings of its learned computation (Kanai, Takatsuki, and Fujisawa 2025), or retrieval-augmented variants that consult a datastore of training instances at inference time to modulate next-token probabilities (Khandelwal et al. 2020). On the epistemological level, however, the core issue would likely remain unchanged: even if such metadata were systematically exploited, it is still unclear in what sense the system itself thereby *takes* these states as reasons on an internalist JTB interpretation, rather than us describing a mechanically implemented pattern in reason-like terms. The temptation to read genuine reasoning off such artifacts is especially visible in current LLM practice: chain-of-thought outputs can function as post-hoc rationalizations and can be systematically unfaithful to the factors actually driving the model's prediction (Turpin et al. 2023; Barez et al. 2025).

their goal was to design an algorithm capable of detecting the features of the human face that are associated with "criminality"—"[...] free from the myriad biases and prejudices that cloud human judgment" (Bergstrom and West 2026). The result of the study was an ANN with an impressive AUC-ROC of 0.95.[6] Confident in that outcome, Wu and Zhang (2016) deduced a "law of normality for faces of non-criminals" based on the results of their classifier.

If we grant, for the sake of argument, that the training procedure was technically sound and that the reported performance is a robust indicator of how the network behaves on that dataset, then, from an internalist (Bayesian) point of view, each classification produced by Wu and Zhang's ANN could be regarded as satisfying the JTB conditions. Precisely this, so the externalist argument goes, shows why a focus on the internal process alone is not sufficient. According to Durán (2025), it is epistemically circular—"it confounds justification of a technically correct output with the justification required for believing that output" (ibid., p. 7). One cannot legitimately justify the ANN's outputs by appealing to the very same functions that produced them: "[...] Wu and Zhang can pin down the functions and properties of their CNN that account for the high predictive accuracy, but at no point can they use those functions and properties alone for claims about justification" (ibid., p. 7). In other words, an internalist account of JTB in this context risks failing to distinguish between what we are compelled to believe by a highly reliable procedure and what, in addition, actually warrants that belief.

According to externalism, the problem with the above approach is that, regardless of the internal coherence and transparency of the ANN, the authors have no justification for believing that someone is a criminal based on their facial traits alone: there are no external factors that tie this hypothesis to any accepted body of scientific knowledge. Accordingly, a proposition stated under that assumption (the classification of the ANN) does not warrant JTB. If we thus switch our focus back to CLERE, the question becomes how we might confer reliability on its outputs in the externalist sense, and, as we have seen in the last section (Section 3.1.2.2), one of the main approaches in that context is process reliabilism. Process reliabilism, as we know, holds that a subject is justified in believing that p if this belief is formed by a process that tends to produce true—or at least truth-conducive—beliefs in a sufficiently high proportion of relevantly similar cases. Applied to algorithmic systems like CLERE, this general idea gives rise to what Durán calls *computational reliabilism* (CR) (ibid., p. 7), an attempt to transfer the justificatory force of reliable computational processes to the outputs of complex simulation and machine-learning models. Instead of asking whether we can extract enough

[6] AUC-ROC—i.e., the area under the *receiver operating characteristic* (ROC) curve—is a number that summarizes how well a classifier can tell two classes apart (here: "criminal" vs. "non-criminal") by measuring how reliably it assigns higher scores to one class than to the other across all possible decision thresholds (Fawcett 2006). Intuitively, an AUC of 0.95 means that in about 95% of randomly chosen pairs consisting of one "criminal" and one "non-criminal" face, the model ranks the "criminal" face as more criminal (gives it the higher score).

internal structure from CLERE to regard its outputs as transparent, CR asks whether the entire algorithmic process—from problem formulation to data collection, model design, evaluation, and deployment—is embedded in a network of what Durán calls *reliability indicators* (Durán 2025, p. 8).

On Durán's account, these reliability indicators fall into three categories: First, there are indicators of technical performance (ibid., p. 8). These concern how the algorithm is specified, coded, trained, validated, and maintained. When they are competently implemented, they support the claim that the algorithm behaves stably and with low error rates across the intended input regimes. Second, there are indicators rooted in computer-based scientific practice (ibid., p. 9). These cover the extent to which the algorithm operationalizes accepted theories, causal models, domain concepts, and evidential standards from the relevant field rather than merely performing decontextualized pattern-finding (as in the example above). Here, expert domain knowledge and knowledge-based integration play a central role: specialists help decide which variables matter, how to encode them, what counts as a meaningful category, and how to relate outputs to existing taxonomies and explanatory frameworks. Third, there are indicators linked to the social construction of reliability (ibid., p. 9): peer review, replication, critical debate, regulatory oversight, and other collective processes through which scientific communities test, contest, and eventually (provisionally)[7] accept or reject algorithmic outputs as JTB.

Summary To summarize the externalist view, we might say the following: Where the internalist perspective asked whether CLERE can itself occupy a standpoint in the "space of reasons"—having beliefs, evidence, and justifications in anything like the human sense—the externalist perspective brackets these questions and treats CLERE as a candidate belief-forming process. On a computational reliabilist reading, CLERE's outputs contribute to knowledge if they are produced by a process that is technically robust, scientifically well-integrated, and socially validated by appropriate communities—that is, if CLERE is embedded in a rich network of reliability indicators. In that case, one could argue that CLERE "knows that p" whenever p is generated by such a reliable computational process operating within its proper domain of application. If, by contrast, CLERE is merely an opaque pattern-matcher optimized for benchmark scores without such external anchoring—what Durán calls "just a bunch of data analysis (JBDA)"[8]—then, like Wu and Zhang's classifier, its outputs do not warrant JTB.

[7] An important consequence of this notion is that CLERE's reliability—and with it, any knowledge attributions grounded in its outputs—is local, fallible, and provisional. Different configurations of reliability indicators will make it dependable in some contexts while leaving it unreliable in others. Like any other scientific proposition, CR does not promise absolute assurances: reliability indicators can change in weight over time, new failure modes can emerge, and previously trusted practices can be revised (Durán 2025, p. 18).

[8] "To my mind, JBDA ignores the 'bigger picture' of knowledge integration, interpretation, and operationalization into algorithms. In fact, I consider JBDA as misleadingly portraying algorithms as

4.1.3. Does CLERE Know That?

From the discussion so far, we draw two conclusions: (1) Without any modification to the classical notion of JTB, the question cannot be applied to a non-human entity such as CLERE without committing a fundamental category mistake and must therefore be rejected. (2) If we widen the terms involved in the definition, there are two main ways to accept CLERE as a knower in the propositional sense: by reconstruing its probabilistic outputs and training dynamics in internalist, quasi-Bayesian terms as belief-like credences with an internal justification, and by adopting an externalist, computational-reliabilist perspective that treats CLERE as part of a reliable belief-forming process whose outputs inherit justificatory status from a surrounding network of reliability indicators. In the first case, we must abandon the notion that the epistemic process is connected to a subject capable of occupying a standpoint in the space of reasons, and in the second case, we effectively relocate the epistemic subject to the collective level, i.e., it is not CLERE that knows, but the wider collective in which it is embedded.[9]

an objective, unambiguous, and scientifically grounded examination of data that produces scientifically meaningful outputs" (Durán 2025, p. 15).

[9] A notion that we will revisit in Section 4.2.2.

4.2. Tacit Knowledge

The second framework we have introduced as a way that CLERE could be said to know is the notion of tacit knowledge (Section 3.2). In contrast to propositional knowledge, tacit knowledge, as we have seen, focuses on aspects of knowing that are manifested in skillful performance, embodied coping, and context-sensitive responsiveness rather than in the explicit endorsement of propositions. Just as in the propositional case, the relevant concepts were originally developed in order to capture what it means for a human subject to know. It is this starting point that has led tacit knowledge, especially in the context of the symbolic AI debate, to be closely associated with authors like Hubert Dreyfus, who treated "know-how" as marking a specifically human dimension of intelligence and as a limit to what computers—understood as formal symbol-manipulating machines—can do (H. L. Dreyfus and S. E. Dreyfus 1992).

Going into the application of this framework to CLERE, we are therefore pulled in two directions. On the one hand, tacitness is traditionally tied to those qualitative features of human agency that seem to distinguish human from machine forms of knowing: embodied skill, holistic situation-understanding, and socially embedded practice. Read in that way, the appeal to tacit knowledge appears to count against the idea that a non-human subject could be said to know anything in this sense. On the other hand, and unlike in the case of propositional knowledge, many tacit-knowledge accounts—as we have introduced them from Ryle to Collins—are not built primarily around explicitly human-specific notions such as "belief" and "justification."[10] Their key concepts do not, at least on some readings, automatically induce a category mistake when we shift the focus from human agents to artificial systems.[11]

Indeed, contemporary machine-learning research explicitly attempts to operationalize Polanyi's paradox ("we know more than we can tell") by building systems that learn from human examples and infer tacit regularities we ourselves cannot fully articulate or program. If we recall the observation from the CLERE section:

[10] Although, as we shall see, tacitness is, on a strict Polanyian reading, anchored in embodied, situation-sensitive coping and thus paradigmatically tied to a human form of life, many of the key notions mobilized in tacit-knowledge accounts are not *per se* (i.e. by definition) restricted to human subjects. Concepts such as *know-how* (as opposed to articulated rules), degrees of *explicability*, the role of *context* and *skill* in competent performance, or Collins's distinction between different *locations* of non-explicitness (RTK/STK/CTK) can, at least in principle, be applied to artificial systems without immediately collapsing into a category mistake. See Héder, Paksi, and The Polanyi Society 2018; Collins 2013.

[11] On these readings, the anthropocentric tilt of the framework is more implicit and historical than definitional: it reflects the paradigmatic human-centered cases and motivations that shaped the discussion, rather than a conceptual constraint that would *in principle* exclude artificial systems like CLERE; see, for example, (Héder, Paksi, and The Polanyi Society 2018; Collins 2013). By contrast, other authors defend an explicitly embodiment-centered interpretation that ties tacit knowledge constitutively to an organic, socially formed human body and therefore denies its extension to AI systems by definition; see, e.g., (Funk 2023).

> [...] whatever we are prepared to call CLERE's "knowledge" is to a significant extent inaccessible to direct inspection: even if the trained network can reliably deploy what it has learned across a variety of tasks, there is no straightforward way to "look inside" the model and recover that knowledge as an explicit set of reasons, rules, or propositions (Section 2.10).

It appears that what we have described with CLERE has a structural profile that is, at least prima facie, closer to the paradigm of tacit knowing than to the paradigm of explicit, propositionally articulable knowledge. CLERE's competence primarily manifests itself in skillful performance and reliable task-sensitive discrimination, while the basis of that competence is encoded in a way that resists extraction into an explicit rulebook or a transparent sequence of reasons. In this respect, CLERE thus seems to exhibit at least a *functional analogue* of tacitness: a practical ability (competence) whose operative organization is real and efficacious, yet not readily available in the form of articulate instructions, either to the system itself or to subjects outside of it. What we thus need to determine with this last section is whether this analogy is merely superficial—resting on a shared kind of opacity—or whether it supports a substantive attribution of tacit knowledge to CLERE.

4.2.1. Polanyi

We have determined that Polanyi's notion of tacit knowing is not exhausted by the fact that some effective know-how is hard to articulate. It is a theory about the *form* of knowing: how a knower relies on certain elements in order to attend to, achieve, or grasp something else, and how this reliance is bound up with personal participation in a practice. In order to do justice to the philosophical depth of Polanyi's notion of tacit knowing, we must therefore unpack his theory along the three interlocking axes we identified: (1) the from-to structure of knowing (the relation between proximal and distal terms), (2) the role of indwelling, that is, embodiment and personal participation in a field of practice, and (3) the status of explicability—i.e., whether, and in what sense, tacit knowing can in principle be rendered explicit or whether it constitutes an irreducibly inexpressible form of knowing.

4.2.1.1. The From-To Structure of Tacit Knowing

Polanyi's central claim is that every act of knowing has a *from-to* structure: we attend *from* a manifold of subsidiarily relied-upon particulars *to* a coherent focal whole. The subsidiaries are not available as an explicit list of reasons or premises; rather, they function as clues that are integrated into a whole. Polanyi argues that subsidiary and focal awareness are, in an

important sense, incompatible: to the extent that particulars function subsidiarily in terms of something else, they cannot simultaneously be grasped as objects of focal attention in their own right. This explains why shifting attention "backward" to the components can dissolve the very meaning achieved in the whole (as illustrated in semantic satiation or in the tightrope example (Section 3.2.1)). Our initial thought of the analogy between this notion and the way CLERE transforms input into outputs is that, here too, a performance that is intelligible at the level of the achieved result depends on a multitude of sub-symbolic elements that are integrated without being available as an explicit list of premises; we will therefore first spell out the elements on which this analogy rests before turning to the points at which it begins to diverge.

The Analogy Hypothesis as a Conjunction of Two Main Claims

First, CLERE's learned competence exhibits a distinctive kind of *non-explicit organization* that invites comparison with Polanyi's subsidiary dimension. CLERE does not represent what it has learned as a list of detachable premises that could be inspected and then cited as the basis of an output. Rather, what supports its task-level performance is distributed across many parameters and intermediate activation patterns in a high-dimensional representational space. In Polanyi's terms, one might describe these internal features as functioning like a manifold of subsidiary particulars: they are not, individually, what the system is "about" in its performance, yet they are causally and functionally relied upon in producing a focal outcome. Importantly, this subsidiary level is not merely hidden from easy inspection; it is also holistic and context-sensitive in its contribution. Individual units or directions in activation space typically do not carry a stable, context-independent content that could be straightforwardly read off as "the" reason for the output. What a given feature does depends on its embedding in a broader activation pattern and on the specific input distribution at hand. The first analogue to Polanyi, then, is that CLERE's competence has a structurally "tacit" profile in the sense that it resists straightforward codification into an explicit rulebook, even when its performance is stable and effective.

Second, CLERE's competence is acquired and manifested in a way that resembles *skill-like performance* rather than the possession of an explicit theory. Polanyi's from-to structure is not a static relation between stored items but a description of an *act of knowing*: a knower integrates subsidiarily relied-upon elements into a focal grasp that can guide judgment and action. Analogously, CLERE's competence is not a mere database of stored associations but a dynamically deployable ability produced through repeated, feedback-driven adjustment. Gradient-based learning reshapes the network across many episodes of training so that it becomes increasingly capable of reliably integrating complex inputs into task-appropriate outputs. In this sense, CLERE can be described as acquiring something like a trained "sensitivity" to

patterns: what it relies on in producing an output is not an explicitly entertained reason, but an organization sedimented through practice, in which certain cues and invariances are selected and weighted because they have proven useful under feedback. The resulting competence is therefore not an articulated account of the domain but an organized disposition to respond successfully under the relevant standards of success (loss, accuracy, calibration, downstream utility, etc.).

The Point of Divergence

Taken together, these two elements express a structural claim: CLERE can be read as a plausible functional analogue of Polanyian tacit integration insofar as it develops non-explicit, distributed, performance-grounded competence through a history of practice (training). But, although this functional analogy is suggestive, it also marks the point at which Polanyi's conception and CLERE's operation begin to diverge. Unlike the human knower, who actively imbues the focal whole with meaning by drawing on personal experience, bodily attunement, and contextual understanding, CLERE's integration is purely mathematical. CLERE has no focal awareness of what it is integrating towards; the "whole" (for example, the concept "cat" in an image) has no meaning for CLERE beyond its role in an optimization problem defined by a loss function and training data. Any semantic content we ascribe to the network's outputs depends on our prior decision to treat a certain label as meaning "cat" and to interpret the mapping as classificatory success.[12]

The analogy between CLERE's complex causal transition from input to output and the from-to structure within tacit knowing would thus likely be rejected by Polanyi if it is read as an analogy *within the system*, because Polanyi's from-to relation is not merely a description of integration but of *meaningful* integration in an act of knowing. Subsidiaries do not function as mere causal antecedents of a result; they function as *clues* whose role is defined by the way they contribute to what is focally meant. As Polanyi puts the point:

> It is their meaning to which our attention is directed. It is in terms of their meaning that they enter into the appearance of that *to* which we are attending *from* them. (Polanyi 1967, p. 12)

In the surrounding discussion, Polanyi's example is the already mentioned case of physiognomic recognition: the features of a face are not merely registered and then summed; rather, one relies on an awareness of them as bearing a joint meaning (a mood, a characteristic expression).

[12] In Section 2.10 we introduced that point in the context of the problem of shortcut learning and dataset bias.

Crucially, this is possible even when one cannot specify the particulars one relied on. The tacitness here is therefore not just opacity about components, but the fact that subsidiaries are indwelt in a way that yields a focal grasp whose "aboutness" is internally constituted by the knower's directed attention and participation. CLERE, by contrast, does not attend to anything as meaningful: its internal features do not show up to it as clues, and its outputs do not present themselves to it as focal wholes that are meant or understood. What occurs inside the network is a sequence of transformations that are normed only by an externally imposed objective, not by an internally lived orientation toward meaning.

This is the sense in which the functional analogy reaches its limit: with regard to meaning, CLERE is in the relevant respect like any other tool. A hammer does not "contain" the meaning of the nail it drives; the meaning arises from the carpenter's purposive use of it within a practice. Likewise, CLERE's "cat"-output is meaningful only within a human practice that (i) fixes the task, (ii) supplies the training signal, and (iii) treats the model's responses as answers to our questions. The argument therefore seems to gain the upper hand that what seems like a Polanyian from-to structure is located less in CLERE's internal processing itself than in the stance of the human users: we choose to regard pixels, embeddings, or intermediate activations as "subsidiary clues" and the output as their "focal" integration, whereas for CLERE there is only a high-dimensional causal dynamics without intrinsic directedness.

4.2.1.2. Indwelling

We have established that unlike Polanyi's human knower, CLERE has no intention in virtue of which the achieved "whole" could be said to be *meant*. In Polanyi's sense, meaning is not an optional semantic label we attach from the outside, but something internally constituted by the knower's directed reliance on subsidiaries as *clues* toward a focal target. The second axis of Polanyi's theory which is connected to that notion, is his concept of *indwelling*: the way in which tacit knowing is rooted in bodily participation and in the incorporation of tools, instruments, and practices into one's mode of "being oriented toward the world."

Polanyi's guiding thought here is that knowing is not merely a matter of having representations but of relying on a lived bodily background from which the world can be attended to and acted upon. As he writes:

> Our body is the ultimate instrument of all our external knowledge, whether intellectual or practical (Polanyi 1967, pp. 15–16).

According to Polanyi, the body is not just an object among others; it is experienced subsidiarily, as that *from* which we attend to things, and it is through skilled use that it is "felt" as one's own—a notion that, as we have seen, Polanyi generalizes to tool use: when a practice is

mastered, the tool ceases to be thematically inspected and becomes a medium through which one is directed toward distal targets (as with a blind man's stick or the scientist's instruments). Indwelling, on this view, names a distinctive kind of incorporation: the proximal term is not merely causally connected to the distal task but is taken up into the agent's lived orientation such that it functions as an extension of the knower's own capacities.

At first glance, this seems to end any analogy. CLERE, as an artificial neural network, does not have a body in Polanyi's sense: it has no lived sensorimotor agency, no proprioception, no vulnerable organismic situatedness, and no practical need that would make a world show up to it as a field of saliences and solicitations. Its physical realization is limited to computational processes instantiated in hardware; even when those processes are coupled to inputs and outputs, the relevant question is whether the system thereby acquires anything like the subsidiary bodily background Polanyi treats as the ultimate instrument of knowledge.

But the deeper point is not only that CLERE lacks a body; it is also that, in the standard case, CLERE does not even encounter the world under the conditions in which indwelling becomes possible. The domain on which CLERE trains is itself already a product of embodied and socially embedded activity. Texts, labels, measurements, classifications, and benchmarks are not neutral givens; they are crystallizations of prior practices of attention, discrimination, and interpretation. In this sense, CLERE does not relate to the world in the way Polanyi's blind man relates to the pavement through his cane; it relates to the world through a residue of human indwelling, namely through data that already reflects what human agents have found salient, nameable, and actionable.

One might object here that this is not a decisive asymmetry, because human knowers also do not encounter the world "as it is in itself." Here it is useful to briefly (re-)locate Polanyi within the broader epistemological landscape introduced in the last section (Section 3.1): We have seen that empiricists such as Hume undermine the picture on which the mind simply "reads" necessary connections off the world: what experience yields, so Hume argued, is at most regular succession (constant conjunction), while the felt necessity involved in causal expectation arises from habit rather than from any perceived tie in the objects. We have also briefly shown Kant's perspective, which holds that even if all our knowledge begins with experience, it does not thereby arise wholly out of experience, because experience itself is possible only under a priori forms and categories that structure what can appear to us as an object at all. On this view, what we encounter is never a raw world unshaped by cognition, but a world as disclosed under conditions of sensibility and understanding—as appearance (Erscheinung), not as thing-in-itself (Ding an sich).

Following this Kantian perspective, the argument that CLERE is only connected to the world through mediating structures—i.e., data—cannot by itself show that it falls short of human

knowing since, so the Kantian argument would have it, humans, too, do not have any epistemic contact with an unconceptualized "given"; what counts as an object of knowledge is already filtered through forms of receptivity, conceptual capacities, and socio-historical practices.

However, and this is the main point, this line of thought does not reinstate the analogy; it rather sharpens what is distinctive about Polanyian indwelling, since, for Polanyi, the relevant contrast is not "direct" versus "mediated" access to reality. Polanyi can happily grant that knowing is always mediated by instruments, skills, and interpretive frameworks. His claim is rather that the mediation is *lived* and *appropriated*: it is part of an agent's practical orientation, sustained by bodily capacities and embedded in a tradition of competent performance. Even if we read Kant as showing that human "givenness" is structured as well, Polanyi's point is that this structured encounter is nonetheless an encounter *for a subject* whose body is implicated in success and failure, whose coping can break down, and for whom the world is disclosed as a field of affordances and risks. CLERE, on the other hand, has no corresponding dimension of vulnerability, concern, or breakdown that would ground the distinction between what is merely processed and what is practically meaningful.

With regard to our hypothesis in question, we can thus conclude that if the from-to analogy breaks at the point where Polanyi requires meaningful integration in an act of knowing, indwelling explains why: the locus of meaning-constitution is the agent who dwells in a practice and is oriented toward a world. CLERE, by contrast, does not interiorize its own operations as a lived medium, and it does not disclose a world as a field of practical significance; it executes transformations whose "success" is defined relative to externally imposed objectives. Whatever epistemic directedness the overall activity has is therefore not an intrinsic feature of CLERE's processing but a feature of the socio-technical arrangement in which humans train, evaluate, and use the model.

4.2.1.3. Explicability

We have reached a point in our discussion where the arguments against our initial hypothesis of analogy between the Polanyian notion of tacit knowledge and an artificial learner such as CLERE have gained an upper hand. The first two axes of his theory have shown that Polanyi's tacit knowing involves meaningful integration in an act of knowing and is anchored in indwelling and personal participation in a practice. CLERE, as we have shown, lacks precisely this internal locus of meaning-constitution and this bodily, commitment-involving mode of orientation toward a world. What remains as the task of this last section is therefore to clarify why the analogy nonetheless suggests itself so readily—i.e., why, despite the evident mismatch, "tacit knowledge" appeared as a promising term to describe CLERE's surprising competence.

The methodological temptation that has emerged throughout the preceding attempt to draw an analogy between Polanyi's tacit knowing and CLERE's learned competence is the tendency to treat every instance of opacity, inexplicitness, or current explanatory failure as evidence for tacit knowledge. We set out this section with the notion that CLERE's learned competence is real and efficacious, yet it is not readily convertible into a compact, surveyable set of rules or reasons, and we therefore formed the hypothesis that this may be described using Polanyi's notion of tacit knowledge. But, if we apply the third axis of Polanyi's theory, we can see that this inference might be too quick unless we first clarify what, exactly, the relevant contrast with the "explicit" is supposed to be.

In their work on *Tacit Knowledge*, Gascoigne and Thornton (2013) propose a way of disambiguating the contrast. They argue that the logical space for tacit knowledge is found by rejecting both the intellectualist principle of codifiability—the notion that all knowledge can be fully stated in context-independent general terms—and the inflationary principle of inarticulacy—that there is knowledge that cannot be articulated at all.[13] According to the middle position they develop, tacit knowledge is uncodifiable in the important sense that it cannot be rendered fully in detached, general instructions, even though it *can* be articulated practically and context-dependently—for instance through demonstration, correction, exemplification, and other situated ways of making one's grasp manifest. The key distinction they introduce is therefore the one between *codification* and *articulation* (Gascoigne and Thornton 2013, pp. 7–8): opacity with respect to the former, so they argue, does not entail absence of the latter.

Applied to the question at hand, this distinction yields an important constraint on how Polanyi's paradox may be operationalized. Although the fact that CLERE's internal organization is not readily codifiable into an explicit rulebook may indeed support a *negative* point—i.e., that its competence is not transparently captured by context-free propositions or reasons—this negative point by itself does not yet license a *positive* attribution of tacit knowledge. For, according to Gascoigne and Thornton, tacit knowledge is not simply "whatever cannot (currently) be made explicit"; it is a form of practical, context-dependent know-how whose content can be made manifest in performance and, in suitable circumstances, articulated in ways that presuppose and address competent uptake (ibid., pp. 191–192).

According to our reading of Polanyi, he as well does not claim that the subsidiarily relied-upon particulars are in principle beyond focus; rather, what he stresses is a kind of "logical" or "weak" unspecifiability: the very integration that constitutes skillful performance and focal

[13] Their guiding aim is, as they put it, "[...] to steer a course between codification and ineffability" (Gascoigne and Thornton 2013, p. 7): tacit knowledge stands opposed to what can be codified in context-independent general terms, "[...] but it does not stand opposed to what can be articulated in any way at all" (ibid., p. 7).

meaning is disrupted when one attempts to render its components as a fully surveyable set of explicit steps. On this reading, the limits of explicability are not the limits of intelligibility *as such*, but the limits of a particular ideal of making knowledge explicit: exhaustive, context-free codification. This also sharpens how we should describe CLERE's opacity. The fact that CLERE's learned competence is not readily translatable into an explicit rulebook may show that it fails the intellectualist demand for codifiable reasons; but it does not by itself indicate that CLERE possesses the kind of tacit knowing Polanyi has in view.

Summary The Polanyian analysis has shown why the initial analogy between CLERE's impressive, non-explicit competence and tacit knowing is tempting, yet ultimately unstable. (1) Along the first axis, the *from-to* structure suggested a functional parallel: CLERE's outputs depend on the integration of many sub-symbolic "clues" that cannot be straightforwardly listed as premises. But the point of divergence was decisive: for Polanyi, subsidiaries are clues only in terms of a focally meant whole, whereas CLERE's internal elements are not integrated in an act of meaningfully directed knowing. (2) Along the second axis, indwelling revealed the deeper ground of this difference: tacit knowing is anchored in embodied participation, vulnerability to success and failure, and the personal appropriation of practices and instruments. CLERE lacks this bodily, commitment-involving mode of orientation toward a world; whatever directedness the overall activity has is supplied by the socio-technical arrangement in which humans train and use the model. And (3) along the third axis, explicability clarified the methodological pitfall: opacity or uncodifiability is not yet tacit knowledge in Polanyi's sense, because Polanyi's claim targets the limits of context-free codification within a practice of skilled, meaningful performance, not merely the current limits of external inspection.

4.2.2. Collins

The assessment of our initial hypothesis that we might attribute the Polanyian concept of "tacit knowing" to CLERE since CLERE's learned competence is real and efficacious, manifests itself primarily in skillful performance, and resists codification into a compact, surveyable set of explicit rules or reasons, has led us to a point where we can conclude that this attribution cannot straightforwardly be substantiated in Polanyi's work itself, since for Polanyi, (i) the from-to relation is not merely an input-output integration but a structure of meaningful reliance; (ii) indwelling ties this meaning-constitution to embodied participation in a practice that is unavailable to CLERE on a fundamental level; and (iii) limits of explicability do not automatically entail a form of tacit knowing. Under this reading, CLERE's competence supports at best a *weaker* claim of tacit-*like* organization (functional tacitness), while the stronger Polanyian notion of tacit *knowing* remains tied to a human subject's lived orientation toward the world.

This does not, however, render the appeal to tacit knowledge pointless; it rather relocates where the interesting questions lie: if anything Polanyian is present in the vicinity of CLERE, it is more plausibly found in how human agents come to rely on, incorporate, and "dwell in" such models as epistemic instruments than in an intrinsic tacit knowing on the part of CLERE itself. This is precisely the point at which Collins's work becomes relevant, since, unlike Polanyi, Collins develops the notion of tacit knowledge not primarily as a phenomenology of *how* a subject integrates subsidiaries into a meaningful focal grasp, but as a sociologically informed analysis of *where* non-explicit competences reside and *why* they resist being made explicit.

It is important to note, however, that this shift in emphasis is not without controversy: as we have already indicated in the last section, Collins's strategy of approaching tacit knowledge via an "antonym"—that is, via its contrast with the "explicit"—has been criticized in particular by Gascoigne and Thornton (2013) for risking a distortion of the phenomenon it aims to analyze. On their reading, Collins's guiding maxim ("explain 'explicit,' then classify tacit") is based on a shift in the concept of explanation: what begins as "not explicated" in the innocuous sense of "not made clear" becomes "not explicable" in the stronger sense of "not explainable" (ibid., p. 8). The result is a shift of attention away from tacit knowledge as a cognitive state or skilled know-how for a subject and toward the "nature of the task, per se" (ibid., p. 9)—something that can be treated as continuous across humans, animals, and even artifacts like sieves. This, they suggest, is why Collins can both deny that cats, dogs, trees, or artifacts "know" anything and yet insist that, apart from deliberative choice, human performances are continuous with theirs: what is doing the explanatory work is no longer the agent's epistemic grasp but an abstract characterization of the task and its causal production (ibid., pp. 9–10).

The direction of this criticism resonates with the constraint that emerged from our Polanyian

analysis and from Gascoigne and Thornton's own distinction between codification and articulation: "tacit" should not be treated as a mere label for whatever is presently opaque, nor should the status of a subject's know-how be determined by whether someone else could, in principle, supply an explicit theory of the task. For the present purpose, however, it is not necessary to refute Collins on this point. Once we have concluded that CLERE does not possess tacit knowing in Polanyi's sense, the productive question is no longer "does CLERE know tacitly?" but rather "what kind of opacity are we dealing with, and where does the epistemic work actually take place?" Collins's taxonomy provides precisely the diagnostic apparatus needed to address this question.

We begin accordingly with Collins's notion of relational tacit knowledge and ask how far the opacity surrounding CLERE can be traced to practice-relative conditions of communication and understanding. Second, we then turn to somatic tacit knowledge, where the guiding question becomes how the material realization of a competence constrains what counts as an intelligible and usable "explicit" description for different kinds of agents. Finally, we take up collective tacit knowledge, which shifts the focus from individual competence to the social and institutional background against which competences are learned, stabilized, and transmitted.

4.2.2.1. Relational and Somatic Tacit Knowledge

Collins's first category, relational tacit knowledge, designates cases in which knowledge remains tacit not because it is in principle ineffable, but because it is not made explicit under prevailing social or practical conditions. Tacitness—or non-explicability—here is therefore due to contingent constraints on disclosure and uptake—for instance concealment, mismatched salience between teacher and learner, dependence on ostension, or the fact that relevant know-how is present but unrecognized. Much of what we have described as inscrutable about CLERE's learning regime and resulting competence can be interpreted in this relational sense.

We have shown, for instance, that mismatched salience is a genuine issue in interpreting machine learning models in the context of shortcut learning or dataset bias (Section 2.10): what the network considers "important" (a certain statistical pattern) might not be obvious or salient to human observers, causing a gap in explication. Likewise, unrecognized knowledge often occurs when a system like CLERE exploits a feature in the data that even domain experts didn't realize was predictive. For example, recent studies have shown that in medical imaging, deep learning systems trained to detect diseases like pneumonia in chest radiographs can achieve high apparent performance partly by latching onto site- and acquisition-specific artifacts such as corner markers, laterality tokens, or other hospital-identifying overlays and imaging signatures—that correlate with where and how the scan was taken rather than with radiographic signs of pneumonia themselves (Zech et al. 2018). The resulting competence is

initially "tacit" in the sense that neither clinicians nor developers recognize it as a basis of the model's discrimination.

In summary and with regard to our question, we can thus state that if the non-explicitness at issue is relational in Collins's sense, then what is missing is not an ineffable inner content but the right conditions for disclosure and uptake: different documentation practices, better dataset and interface transparency, targeted audits and interpretability work, or simply the removal of incentives to conceal. In this respect, RTK supports a comparatively deflationary reading of CLERE's opacity: at least some of what appears inscrutable may be rendered explicit (or at least communicable in publicly checkable ways) once we know what the system is exploiting and how this reliance is stabilized by the surrounding training and evaluation regime.

The second category that Collins introduces, somatic tacit knowledge, refers to tacit knowledge grounded in the human body and nervous system. We have seen that Collins is pursuing two objectives here: on the one hand, he wants to show, in continuity with the other categories, what makes the bodily "knowledge" and "skill" non-explicit in a tacit sense, and on the other hand, he is determined to "demystify" the notion of the human body as something epistemologically special. In order to do so, he divides STK into two subtypes: somatic-limit tacit knowledge, which refers to cases where knowledge can be explicitly formulated but humans cannot use the explicit instructions due to our physiological or cognitive limits (we are too slow or the information is too fine-grained for conscious processing), and somatic-affordance tacit knowledge, where knowledge is not explicit simply because our bodies handle the task automatically (we have no need for an explicit formula).

The original argument that Collins puts forward here is thus directly opposed to what we have discussed in the context of the Polanyian notion of embodiment and indwelling. For Polanyi, embodiment is epistemically relevant because it anchors a distinctive mode of "being in the world": bodily vulnerability, practical involvement, and a lived from-to orientation are not merely enabling conditions for performance, but conditions for meaning-constitution and hence for tacit knowing in the strong sense. Collins's point, by contrast, is methodologically deflationary. The fact that CLERE does not possess an organic body is, for Collins, not primarily an objection that would locate a principled epistemic gap between human and machine. Rather, it is a question about *how* a competence is causally realized and about which materials and temporal regimes make certain realizations feasible.

Following this line of argumentation, the fact that CLERE lacks a body is primarily an empirical and engineering constraint: the question of whether a given competence can be made explicit or mechanized is not "can this ever be formulated?" but rather "what material and temporal regime would make an explicit or mechanized realization workable?" In Collins's own examples, bodily skills become "tacit" not because they hide a mysterious inner content, but because the

relevant information is either too fine-grained, too fast, or too densely coupled to the dynamics of the organism to be followed by conscious, stepwise rule application.

Collins's distinction between somatic-limit and somatic-affordance thus provides two answers to the question of why CLERE's learned ability often resists explication: First, the somatic-limit analogue concerns limitations of *use* rather than of *formulation*. In principle, one can always "make explicit" what CLERE does by presenting the trained parameters, the architecture, and the full forward pass as an executable specification. But this kind of explicitness is not the kind that supports understanding and transfer for human agents: it is explicit only in a machine-readable sense. The obstacle is therefore not ineffability but tractability and uptake: the functional description is too high-dimensional, too uncompressed, and too temporally fine-grained to be employed as a guide for human reasoning or instruction.

Second, the somatic-affordance analogue highlights the way in which competences can be carried by automatic, subpersonal routines that never need to be rendered as explicit rules at all. For humans, this is the familiar sense in which we balance, track, or coordinate movement without consulting a formula. For CLERE, something similar holds at the level of learned pattern sensitivity: once training has stabilized a reliable discriminative routine, the system has no operational need to represent its own grounds as a surveyable set of reasons. The competence is "there" in the disposition to respond correctly across contexts, but it is not stored as an explicit inventory of rules, and it is not accessed by anything like deliberative endorsement. On this view, the non-explicitness is a by-product of an efficient division of labor between training (where structure is acquired) and inference (where it is deployed automatically).

In summary and with regard to our guiding question, STK thus supports a second, equally deflationary diagnosis of CLERE's "tacitness": the obstacle is often not formulation but use. CLERE's competence can be made explicit in a strict, machine-readable sense (architecture, parameters, executable forward pass), yet this explicitness typically fails to be human-tractable and thus does not function as an epistemic resource for explanation or instruction. Conversely, once training has stabilized a reliable routine, there is no operational pressure for the system to store its grounds as a surveyable set of rules or reasons; non-explicitness is, in this respect, a by-product of how the competence is materially realized and automatically deployed.

4.2.2.2. Collective Tacit Knowledge

With the third and last category—*collective tacit knowledge*—Collins introduces the strongest and, for our purposes, most consequential sense in which "tacit knowledge" exceeds what can be rendered fully explicit. In CTK, non-explicit competence is not merely a matter of contingent non-disclosure (as in RTK) or of bodily realization (as in STK), but of social embedding: the relevant know-how is sustained by a community's practices, languages, standards, and role-distributions in such a way that it cannot be exhaustively "possessed" or articulated by any individual taken in isolation. With regard to the question at hand, let us therefore consider what Collins called the *socialization problem* (Section 3.2.2.3): although we can describe the circumstances under which CTK is acquired, we lack a concrete specification (codification) of the mechanism by which an individual remains "in touch" with a collective and its continually changing body of knowledge.

Although Collins does not explicitly differentiate between symbolic and subsymbolic AI,[14] his formulation of the socialization problem was clearly developed in the context of symbolic AI systems—the kind of rule-based, lookup-table approaches exemplified by Searle's Chinese Room he references. His central argument is that such systems face an insurmountable obstacle: maintaining fluency with a living language or social practice requires continuous manual updating by a human attendant who recognizes changing contexts and conventions. As Collins puts it: "It takes a human to socialize the machine because we have not yet solved the socialization problem" (Collins 2013, p. 130). The problem, on this view, is that language and social norms change continually—new idioms emerge, old terms become archaic, contextual proprieties shift—and a rule-based system operating on a frozen database will soon become "noticeably out of date" (ibid., p. 130). Even if an attendant continually updates the system's rule base, "[...] all the social intelligence would be in the attendant, [...]" and the machine would remain, epistemically speaking, "[...] the attendant's marionette" (ibid., p. 130).

A critical voice might object, however, that this diagnosis rests on assumptions about AI architecture that no longer hold in the era of large-scale machine learning. CLERE, after all, is not a symbolic lookup table but a subsymbolic learner: it is trained on vast corpora of socially produced linguistic practice, and it can be retrained, fine-tuned, or continually updated from new data streams without requiring a human to manually encode new rules or contextual proprieties. One might argue, then, that CLERE provides something like an engineered surrogate for socialization—a mechanism that maintains ongoing alignment with the changing

[14] A conflation that persists throughout the book: although Collins explicitly mentions neural nets as "trained" systems, he does not treat the symbolic/subsymbolic distinction as a structurally relevant axis of his argument. Consequently, it remains unclear whether the socialization problem is intended as a critique of classical rule-based AI in particular or of "AI" in general, including connectionist architectures.

collective and thereby solves, or at least dissolves, Collins's socialization problem. While CLERE does not "live" within a form of life in the ordinary sense, it can nevertheless be coupled to the collective in a technologically mediated way: it can incorporate large-scale feedback signals (human ratings, interaction traces, institutional guardrails); it can be re-aligned at scale whenever ambient norms drift; and in principle, it could even operate in a continual learning regime where the model updates itself incrementally as new data arrives, without requiring a human gatekeeper to decide what counts as a relevant change.[15] In other words, one might hypothesize that CLERE could remain "in touch" with the collective not through embodied participation and social membership, but through systematic, data-driven re-attunement that keeps its dispositions synchronized with the current state of public practice—thereby removing the human from the loop and solving the socialization problem in a way that symbolic AI never could.

If we read Collins's account of the socialization problem more closely, however, this line of argumentation faces two decisive problems. First, we have already determined that the data that CLERE operates on—the linguistic and behavioral traces through which it could be said to stay "in touch" with a collective—is not a neutral "given," but a product of selection, filtration, and institutional shaping. What reaches the model as "the social" is already formatted by platforms, incentives, moderation regimes, editorial decisions, and uneven patterns of visibility; it is a curated residue of practice rather than practice itself. Thus, even if we grant that continual learning techniques could allow CLERE to update itself automatically without a human attendant manually re-encoding rules, the fundamental point remains: the data stream itself is mediated, and someone—some institutional actor, some platform designer, some moderation team—still occupies the role of determining what counts as the collective practice worthy of incorporation. The human has not been removed from the loop; the locus of human judgment has simply shifted from rule encoding to data curation and filtering. Moreover, even if the corpus were a perfectly comprehensive record of public utterances, CLERE would still stand in the wrong relation to the collective whose traces it consumes in the sense relevant to Collins's CTK. It would register outputs without occupying a place in the normative economy that makes those outputs *answerable*: it is not a participant who can be addressed, corrected, sanctioned, trusted, distrusted, promoted, excluded, or held responsible in the ordinary ways that structure membership in a community.

[15] Contemporary approaches to continual or lifelong learning in neural networks aim to enable models to learn from streaming data without catastrophic forgetting of prior knowledge. Techniques include elastic weight consolidation (Kirkpatrick et al. 2017), experience replay (Rolnick et al. 2019), and dynamic architecture expansion (Rusu et al. 2016). While these methods remain active research areas and face significant technical challenges, they suggest that in principle, a system like CLERE could update its own parameters in response to distributional shifts without requiring manual human re-curation of training data. See Parisi et al. (2019) for an overview.

Second, even if we set aside this mediation problem and assume, for the sake of argument, that CLERE had frictionless access to an unfiltered stream of collective practice and could update itself automatically in response to every shift in usage, the core difficulty would remain: socialization is not exhausted by exposure to regularities, however rich or dynamically updated, because what is at issue is precisely the capacity to stay oriented within a *living* space of norms. Collins illustrates this point with the thought experiment of trying to learn Chinese in a purely consultative way: imagine a learner who relies on an exhaustive dictionary and rulebook and therefore has to look up each word (and, in effect, each use) every time anew. Such a procedure may generate locally correct outputs, but it never reaches the kind of practical fluency in which the language is "one's own"—not because the database is incomplete, but because the competence remains parasitic on explicit consultation rather than being sustained by a socially formed responsiveness. The analogy carries over even to a continually learning CLERE: a snapshot of linguistic regularities can always be taken—and with continual learning, these snapshots can be taken frequently and automatically—but what is missing in such a snapshot is precisely what makes those regularities the living language: their continuous renegotiation through usage, the drift of meanings, the emergence of new idioms, the obsolescence of old ones, and the shifting standards of when an utterance counts as apt, odd, rude, humorous, or simply wrong. In that sense, even a perpetually self-updating CLERE would remain "static" in the way that matters for CTK. At any given moment, its competence would be a *fit-to-date*: a model trained (or fine-tuned) on a temporally indexed residue of practice, rather than an active participant in the ongoing, situated renegotiation of what the practice *is*.

What CTK brings into view, then, is not merely a further "hard case" for explicability, but a deeper methodological shift we have already hinted at in the transition from Polanyi to Collins: the explanatory focus moves from the internal structure of an individual act of knowing to the socio-technical conditions under which competent performance is learned, stabilized, recognized, and transmitted. In this sense, Collins's notion of CTK suggests that the individual may be the wrong unit of analysis for at least some forms of knowledge: what is "known" is, in an important respect, sustained between people—across roles, routines, and institutions—rather than inside any single head (or model).[16]

[16]This collectivizing notion connects to a broader debate in contemporary social epistemology: how epistemic achievements depend on social coordination, norms of trust and authority, divisions of cognitive labor, and institutional mechanisms that certify and validate what counts as knowledge (O'Connor, Goldberg, and A. Goldman 2024; A. I. Goldman 1999). Knowledge—as we have discussed it throughout this thesis—has typically had a strongly individualistic orientation. Social epistemology, in contrast, understands the formation of knowing not in terms of what occurs inside an individual but in terms of the broader social processes that make knowing possible: testimony, expertise, division of labor, and the role of shared norms in stabilizing epistemic outcomes (Gramelsberger and Sigwart 2025, p. 210).

Collins himself, as we have seen, sharpens this reorientation by diagnosing what he calls *social Cartesianism* (Section 3.2.2.3): the assumption that the primary bearer of knowledge is an individual

Summary

Collins's taxonomy enables a fine-grained diagnosis of the kinds of non-explicitness at issue in CLERE's competence. Whereas Polanyi ties tacit *knowing* to a subject's embodied indwelling and meaning-constituting from–to orientation, Collins relocates the explanatory burden from the phenomenology of an individual knower to the socio-technical conditions under which competences are learned, stabilized, and rendered communicable. Relational tacit knowledge supports a deflationary reading of at least some opacity surrounding CLERE: what appears inscrutable may reflect contingent limits of disclosure, salience, documentation, or auditability rather than an ineffable inner content. Somatic tacit knowledge sharpens a second deflationary point: even when a competence is "explicit" in a strict machine-readable sense (parameters, architecture, executable forward pass), it may fail to be usable as an epistemic resource for human understanding and instruction, while the system itself has no operational need to store its grounds as surveyable rules or reasons. Collective tacit knowledge then marks the strongest limit, shifting the unit of analysis from the individual system to the collective practices and institutions that sustain what counts as competent performance. On this view, Collins's socialization problem is not straightforwardly dissolved by subsymbolic learning or continual updating, since the relevant "contact with the collective" remains mediated by selection and governance of data streams and, more fundamentally, by a missing relation of normative participation and answerability. The upshot is that Collins's framework does not vindicate a strong attribution of tacit *knowing* to CLERE; instead, it helps locate where different kinds of opacity originate and suggests that, for some epistemic phenomena, what is "known" is maintained *between* agents, roles, and institutions rather than *inside* any single head (or model).

subject, and that the social world enters merely as an external source of inputs or optional context. According to Collins, this notion is not simply a philosophical mistake but a deeply rooted methodological reflex that makes us look for the essence of knowing *inside* the agent, thereby overlooking the ways in which competence is maintained by participation in a collective (Collins 2013, p. 131).

4.2.3. Does CLERE Know How?

The application of the tacit knowledge theories lets us draw four conclusions with respect to the question of whether "CLERE knows how":

(i) If "S knows how to φ" is read in the thin, performance-based sense of dispositional competence—i.e., as the stable ability to produce successful outcomes under appropriate conditions without requiring reflective access, explicit rules, or propositional endorsement—then CLERE qualifies in a straightforward way within its trained domain: it has acquired reliable task competence, manifested in context-sensitive, repeatable success relative to the relevant standards of evaluation (loss, accuracy, calibration, downstream utility). However, what CLERE "knows how" to do is fixed by its training distribution, objective function, and deployment regime rather than by a general practical mastery.

(ii) If, by contrast, "knowing how" is taken in the thicker Polanyian sense in which competent performance is the manifestation of a meaningfully directed act of knowing—structured by from-to integration and anchored in indwelling, bodily participation, and a lived orientation toward a world—then CLERE does not qualify: the decisive obstacle is not merely that its operative organization resists codification, but that Polanyian tacit knowing is not exhausted by non-explicitness. Subsidiaries function as clues only within a focal grasp that is meant by a knower, and indwelling locates this meaning-constitution in embodied participation and vulnerability to success and failure; CLERE's integration, by contrast, is a mathematically implemented sensitivity normed by externally imposed objectives, while whatever "aboutness" its outputs have is conferred by the socio-technical practice that fixes tasks, labels, and standards of correctness.

(iii) Collins's taxonomy sharpens what is and is not warranted by CLERE's non-explicitness: much of what makes CLERE appear "tacit" is plausibly traced to relational and somatic sources of non-explicitness, where what is missing is contingent disclosure and uptake (RTK) or where what is explicit in a machine-readable sense is unusable for humans as a guide for understanding and instruction due to material and temporal constraints (STK). On this diagnosis, opacity blocks a demand for compact, surveyable rulebooks, but it does not by itself license a robust attribution of tacit knowing.

(iv) Finally, if the relevant notion of "knowing how" is the collective one made salient by CTK and the socialization problem—competence sustained by a community's evolving practices, languages, and norms in a way that cannot be exhaustively possessed by an isolated individual—then the epistemic subject is best relocated from CLERE to the socio-technical practice in which it is trained, evaluated, corrected, and deployed.

4.3. Summary

This chapter applies the epistemological frameworks developed in Chapter 3 to assess CLERE's epistemic status systematically. The analysis proceeds in two parts: propositional knowledge (JTB framework) and tacit knowledge (Polanyi and Collins).

Propositional Knowledge The classical JTB schema cannot be applied to CLERE without modification, as "belief" and "justification" presuppose a conscious human subject occupying a doxastic perspective. Two post-Gettier strategies offer pathways: (1) An *internalist* approach reconstructs CLERE's outputs as graded credences (Bayesian probabilities) and its training as a disciplined updating procedure, yielding a quasi-justification grounded in loss minimization and learned representations. However, this account relocates epistemic assessment from a personal-level "space of reasons" to a subpersonal cascade of activations, raising the question whether such modified terms still capture genuine knowledge or merely describe reason-like mechanisms. Epistemic classifiers operationalize this by labeling outputs as "I know," "I may know," or "I don't know" based on local evidential support in the training distribution. (2) An *externalist* approach, exemplified by computational reliabilism (Durán), treats CLERE as a candidate belief-forming process rather than a believer. On this view, CLERE's outputs warrant JTB when embedded in a network of reliability indicators: technical robustness, scientific integration with domain knowledge, and social validation through peer review and institutional oversight. The Wu-Zhang criminality classifier illustrates the failure mode: high internal accuracy without external anchoring yields epistemically circular outputs that confound technical correctness with genuine justification.

Tacit Knowledge The Polanyian analysis proceeds along three axes: (1) The *from-to structure* initially suggests an analogy—CLERE integrates distributed, subsymbolic "clues" into task-level outputs without surveyable premises. But the divergence is decisive: for Polanyi, subsidiaries function as clues only within a focally *meant* whole, whereas CLERE's integration is purely mathematical, normed by externally imposed objectives. (2) *Indwelling* reveals the deeper asymmetry: tacit knowing is anchored in embodied participation, vulnerability, and a lived orientation toward a world that discloses practical significance. CLERE lacks this commitment-involving mode of being; its directedness is supplied by the socio-technical arrangement. (3) *Explicability*: Gascoigne and Thornton's distinction between codification and articulation clarifies that opacity alone does not entail tacit knowing—Polanyi's claim targets the limits of context-free rulebooks within skilled performance, not mere current inscrutability.

Collins's taxonomy relocates the analysis from individual phenomenology to socio-technical conditions. *Relational tacit knowledge* (RTK) shows that much opacity stems from contingent

disclosure failures (concealment, mismatched salience, unrecognized patterns like shortcut learning). *Somatic tacit knowledge* (STK) diagnoses a mismatch between machine-readable explicitness (parameters, architecture) and human-tractable instruction, while highlighting that CLERE's competence, once trained, operates automatically without needing surveyable reasons. *Collective tacit knowledge* (CTK) marks the strongest limit: Collins's socialization problem—staying attuned to a living, norm-governed practice—is not dissolved by continual learning, since (i) data streams are curated residues, not practice itself, and (ii) CLERE is not answerable within the normative economy that structures community membership. The upshot: epistemic competence is better located in the socio-technical practice (training, evaluation, deployment, governance) than in CLERE as an isolated system.

Question from Theory Chapter	Conclusion from Application Chapter
Propositional knowledge (JTB)	
Can the JTB schema be applied to CLERE? If so, what could "belief" and "justification" plausibly amount to for a non-human entity, and which parts of the classical package must be revised (if any)?	Not directly—belief and justification presuppose a doxastic perspective. Two modified routes: (1) Internalist: outputs as credences, training as rational updating, justification via loss minimization (relocates to subpersonal level); (2) Externalist: CLERE as reliable process, outputs warrant JTB when embedded in reliability indicators (technical robustness, scientific integration, social validation).
Given that pre-Gettier JTB is incomplete, which post-Gettier strategy is most appropriate for evaluating CLERE's propositional status: an internalist reconstruction of belief/justification, an externalist assessment of reliability (truth-conduciveness), or both as complementary lenses?	Both as complementary lenses. Internalist: probabilistic outputs as credences with internal justification via training; operationalized in epistemic classifiers (IK/IMK/IDK). Limitation: no standpoint in "space of reasons." Externalist: outputs inherit justification from reliability indicators. Wu-Zhang classifier illustrates failure: internal accuracy without external anchoring yields epistemic circularity.
Is there an internalist reading of CLERE's outputs that treats them as belief-like states (e.g. graded confidence) and ties them to an internal notion of justification?	Yes, via Bayesian regimentation: outputs as conditional probabilities, training as updating via gradient descent. Epistemic classifiers operationalize with IK/IMK/IDK labels based on local support. Critical limitation: entirely subpersonal—no agent takes states as reasons. Unclear if this preserves internalist JTB or merely describes reason-like mechanisms.
If CLERE is approached as a belief-forming process rather than as a reason-giving subject, what would count as sufficient reliability for knowledge attribution? Which features of CLERE's development and use would have to anchor its outputs in the world in a truth-conducive way?	Computational Reliabilism: outputs warrant JTB when embedded in reliability indicators across: (1) technical performance (stable, low-error behavior); (2) scientific integration (operationalizes accepted theories, expert-guided variable selection); (3) social validation (peer review, replication, oversight). Epistemic subject relocates to collective level—the socio-technical practice, not CLERE alone.
Tacit knowledge ("knowing how")	
How does Ryle's knowing-how/knowing-that distinction reorient the discussion from doxastic states to competence and thereby set the stage for treating CLERE's "tacit" dimensions primarily via the Polanyi framework?	Licenses focus on performance-based competence rather than propositional endorsement, making tacit knowledge frameworks applicable. However, structural analogies (non-explicit organization, skill-like performance) do not automatically warrant strong tacit knowing attributions—opacity alone insufficient.
Does CLERE's non-explicit organization support more than a merely superficial analogy to Polanyian tacit knowing? How does the analogy look when examined along Polanyi's three axes of (i) from–to integration, (ii) indwelling, and (iii) explicability?	No, breaks on all three axes: (1) From-to: mathematically integrates subsidiaries but lacks meaningful focal grasp—no intrinsic aboutness. (2) Indwelling: no embodied participation, vulnerability, or lived orientation; directedness supplied externally. (3) Explicability: opacity $\neq$ tacit knowing; Polanyi concerns limits of context-free codification in skilled performance. Polanyian tacit knowing appears in how humans incorporate CLERE as epistemic instrument.
Which kinds of non-explicitness are at issue in CLERE's case: relational (disclosure/uptake), somatic (material and temporal constraints on usable explicitness), or collective (dependence on social embedding)? In particular, how (if at all) does Collins's socialization problem bear on CLERE and motivate a shift toward social epistemology?	All three apply: (1) RTK: opacity from contingent disclosure failures (concealment, mismatched salience, unrecognized patterns)—deflationary. (2) STK: mismatch between machine-readable and human-tractable explicitness; trained competence operates automatically without surveyable reasons. (3) CTK: strongest constraint—socialization not dissolved by continual learning: data is curated residue, not practice; CLERE not answerable within normative economy. Epistemic competence located in socio-technical practice, not CLERE as isolated system.

Table 4.1.: Summary of Application Chapter: Conclusions on CLERE's Epistemic Status

5. Conclusion

This thesis set out to address a question that has become increasingly urgent as artificial neural networks achieve remarkable linguistic and inferential capacities: *can systems like CLERE meaningfully be said to know?* The inquiry was motivated by a conceptual puzzle. On the one hand, contemporary ANNs produce outputs that would, in human contexts, manifest epistemic competence—they classify, diagnose, translate, explain, and argue with a fluency that rivals domain experts. On the other hand, the epistemic frameworks through which we have traditionally understood knowledge were developed against the background of human subjects and appear to resist straightforward application to subsymbolic learners whose competence is distributed across high-dimensional parameter spaces rather than articulated in the "space of reasons."

Our approach has been to take this tension seriously—neither dismissing machine competence as mere simulation nor uncritically extending epistemic predicates to systems whose internal organization differs radically from human cognition. The methodological commitment underlying the analysis was epistemological pluralism: the recognition that there is no single, privileged framework exhausting what it means to know, and that different epistemic vocabularies illuminate different aspects of cognitive achievement. Accordingly, the thesis proceeded through three stages: a technical reconstruction of how CLERE achieves its competence (Chapter 2), a systematic development of the relevant epistemological frameworks (Chapter 3), and an integrative assessment applying those frameworks to CLERE (Chapter 4).

5.1. Summary of Findings

The technical analysis established that CLERE's competence arises through a distinctive combination of mechanisms: error-driven self-modification via gradient-based learning, distributed internal representations that resist extraction into explicit rules, and high-dimensional encoding in which downstream tasks become tractable through learned geometric structure. CLERE does not store knowledge as an inventory of propositions nor execute predetermined inference rules. Rather, it develops sensitivity to statistical regularities through iterative exposure to data, compressing what is predictive while discarding noise, and deploying this learned organization automatically at inference time. This profile suggested a functional analogue of tacit knowing: impressive, task-appropriate performance whose operative basis is not readily surveyable as a compact rulebook. For a compact specification of these modeling assumptions, the reader is referred to Section 2.10.

The epistemological analysis then developed two complementary vocabularies. For propositional knowledge, we reconstructed the JTB schema and its post-Gettier refinements, tracing how the justification condition splits into internalist accounts (justification fixed by the subject's epistemic perspective) and externalist accounts (justification requiring truth-conductive processes appropriately connected to the world). For tacit knowledge, we examined Polanyian tacit knowing—anchored in the from-to structure, indwelling, and embodied participation—and Collins's sociologically informed taxonomy distinguishing relational, somatic, and collective sources of non-explicitness. Table 3.1 consolidates the epistemic frameworks developed in Chapter 3 together with the guiding questions each raises for the assessment of CLERE; readers seeking a compact overview of the analytical vocabulary and its operationalization are directed there.

The application showed that attributing propositional knowledge to CLERE requires either an internalist translation into graded credences plus local evidential checks (Section 4.1.1) or an externalist, reliability-based embedding in technical, scientific, and social validation (the "Externalist View" section in Chapter 4); the Wu–Zhang case study illustrated how strong internal performance without external anchoring collapses into epistemic circularity. On the tacit side, we showed that Polanyi's criteria do not transfer straightforwardly: CLERE's integrations are mathematically optimized rather than intentionally meant, ungrounded in embodied participation, and mere opacity is not yet tacit knowing (Section 4.2.1). Collins's taxonomy clarifies that much non-explicitness is relational or somatic, but the strongest limit remains collective tacitness—the socialization problem of staying answerable within a living, norm-governed practice is not dissolved by continual learning from curated data residues (Section 4.2.2). Table 4.1 synthesizes these results into a compact verdict on CLERE's epistemic status across each framework; for the direct conclusions on whether and under what conditions knowledge attributions succeed, the reader is referred to that summary.

5.2. Reflection

In Section 1.2, we introduced what Abel called *knowledge research*: the task of describing the distinctive profiles of various modes of knowing, identifying their points of overlap, elucidating the mechanisms by which they interact and interpenetrate, and grasping the dynamic character of these interactions as they manifest in perception, language, thought, and action. Although, as we have established in our methodology, the focus of this thesis was not knowledge research in Abel's full sense, we now take a step back and note our observations on that level.

The first such observation concerns the relationship between the two frameworks we employed. Throughout the application, we found that neither propositional nor tacit knowledge frameworks,

applied in isolation, could yield an adequate verdict on CLERE's epistemic status. Where the propositional register captured what is at stake in reliability and calibration, it systematically left out the practice-bound dimensions of competence; where the tacit register illuminated the non-explicitness of learned representations, it risked collapsing into a mere redescription of opacity. This is not an accidental result of the particular case but a reflection of what Abel identified as the irreducible plurality of knowledge forms: in any concrete epistemic episode, different modes of knowing are always already fused, and their analytic separation is a heuristic that must be balanced against their constitutive entanglement (Abel and Conant 2012, p. 33). What the CLERE case makes vivid, then, is not only where and why knowledge attributions succeed or fail but something about the structure of knowledge itself—namely, that the conditions of propositional success and tacit competence are not independently satisfiable. It is only the combined use of these frameworks that allows us to say, with greater precision, what is missing when performance is not enough, where the relevant epistemic standing plausibly resides, and which varieties of non-explicitness matter for knowledge attribution and which do not.

If one speaks only in terms of explicit propositional knowledge, the debate is pulled toward questions of output correctness, calibration, and reliability: whether the system gets the answer right, whether its confidence matches its accuracy, and whether it succeeds across test distributions. Although these questions are important, they also invite an implicit shift in what is taken to matter. They encourage us to treat epistemic standing as something that can be read off from performance profiles alone, and in doing so they leave out the practice-shaped dimensions of competence that matter when we ask who (or what) is genuinely a knower—dimensions such as answerability, responsiveness to reasons, and participation in norm-governed correction. If, conversely, one appeals too quickly to "tacit knowledge" as a placeholder for whatever is opaque, high-dimensional, or difficult to reverse engineer in neural networks, one risks an opposite distortion. We have shown that opacity by itself is not yet tacit knowing. It is first and foremost a technical fact about representation and access, and it becomes philosophically significant only under further conditions: that is, when the non-explicitness is linked to a form of situated skill, to embodied or socially cultivated attunement, or to a standing in a practice in which error can be owned, repaired, and normatively assessed. Without those further conditions, invoking tacit knowledge may merely rename the explanatory gap while smuggling in an unwarranted conclusion about epistemic subjecthood.

A second observation concerns the bidirectionality of illumination. We have seen that when epistemological frameworks are applied to a genuinely novel epistemic object such as CLERE,[1] the resulting insights are not simply unidirectional. The frameworks are not merely tools that we bring to bear on an independent subject matter and then lay aside once the "verdict"

[1]Novel from the perspective of the epistemological frameworks brought to bear on it.

is in; the application also serves as a diagnostic for the frameworks themselves: Internalist accounts of justification, when pressed to accommodate CLERE, revealed that their core notion of the "epistemic perspective of the subject" carries more theoretical weight than is typically acknowledged—the assumption of a unified first-person standpoint from which evidence is assessed is not a neutral or minimal commitment but one that presupposes precisely the kind of agential organization that CLERE lacks. Reliabilism, by contrast, proved more accommodating—but the accommodation came at a cost: the relevant process must be extended so far into the surrounding sociotechnical system that the boundary between the system as the locus of knowledge and the system as a component of a larger epistemic practice becomes difficult to maintain. Polanyi's framework, similarly, revealed its tacit anthropological commitments only under some pressure: the from-to structure, indwelling, and embodied participation are not incidental features of his account but load-bearing elements that resist separation from the human form of life in which they were developed. Collins's taxonomy, by contrast, proved more adaptable precisely because it already theorizes tacit knowledge in relation to social structures rather than individual subjects. This diagnostic dimension of our analysis is not just a side effect but a philosophical result in its own right: it suggests that the categories through which we assess machine knowledge are themselves in need of revision as the range of epistemic objects to which they are applied expands.

A third observation, which follows from the foregoing, concerns the location of epistemic standing. One of the most consistent findings of our application is that the epistemic predicates most plausibly attach not to the isolated trained model but to the broader sociotechnical practice in which the model is embedded: the data selection, validation protocols, institutional accountability structures, and forms of ongoing human oversight through which model outputs acquire or fail to acquire genuine warrant. This result might initially appear merely technical—a point about where to draw the system boundary. But it carries a deeper implication. If epistemic standing is constitutively distributed across a practice rather than localized in an artifact, then the customary intuition that knowledge has a subject—a locus to which it belongs and from which it can be attributed—requires reconsideration. The CLERE case does not merely complicate the question of whether this particular system knows; it puts pressure on the assumption, common to both folk epistemology[2] and most formal treatments, that knowledge is the kind of thing that can, in principle, be localized. This points toward what might be called an *ecological* reconception of epistemic standing—one in which the relevant unit of analysis is not the knower but the practice, and in which individual contributions to that practice, whether human or artificial, are assessed not in isolation but in terms of their role within a structured whole. Developing such a conception rigorously would require resources beyond those deployed here, but the need for it is one of the clearest theoretical upshots of our investigation.[3]

[2]The implicit epistemic frameworks embedded in common sense and ordinary language. See Gerken 2017.

[3]The idea that epistemology should take practices and communities rather than individual knowers as

5.3. Further Questions

Having established with some precision what CLERE's epistemic status is—and, equally important, what it is not—we are now in a position to raise a series of questions that follow the epistemological groundwork laid out with our thesis. These questions extend the subject matter beyond our area of focus into the domains of ethics, politics, and the philosophy of knowledge creation. We close by sketching five such questions as invitations for further research.

(1) We have shown that CLERE's competence, however impressive, does not amount to the kind of knowledgeable participation in a practice that characterizes human epistemic agents. This finding has immediate consequences for the ethics of AI and, in particular, for the diversity of knowledge forms that ethical reflection must accommodate. Funk has argued that AI ethics is addressed to embodied, sensory, emotional beings living in communicative communities—not to mathematically reduced mass points or vectors in computational models—and that the reduction of ethical reflection to what machines can simulate threatens to undermine the very "knowledge diversity" (*Wissensdiversität*) on which democratic societies depend (Funk 2024a). What he calls "personal knowledge" in Polanyi's sense—the tacit, interpersonal, tradition-bound competences through which moral communities orient themselves—constitutes, in his analysis, a critical infrastructure that cannot be replaced by formalized guidelines or algorithmic proxies (ibid.). If our epistemological analysis is correct that CLERE lacks precisely this dimension of situated, norm-governed knowing, the question arises:

> How should the deployment of AI systems in ethically sensitive domains be designed so as to preserve rather than erode the diversity of personal, tacit, and practice-based knowledge forms on which moral judgment depends?

(2) A related but more immediately practical question concerns the regulatory frameworks through which societies govern AI. Funk has shown, in his analysis of the EU AI Act and comparable top-down governance instruments, that the reduction of ethics to guidelines—what he terms "checklist ethics" (*Checklistenethik*)—risks treating moral responsibility as a task

---

its primary unit of analysis has been developed along several lines. Longino argues that normative epistemic judgments attach to epistemic communities and their procedures rather than to individuals, and identifies responsiveness to criticism and shared standards of inquiry as the conditions under which community-level outputs earn the status of knowledge (Longino 2002). Code proposes what she calls "ecological thinking" as a restructuring of epistemology around situated, interdependent epistemic agents embedded in social and material environments, explicitly criticizing the "abstract individualism" of mainstream epistemology (Code 2006). Most recently, Elgin has developed a fully articulated "epistemic ecology" in which knowledge is understood as the achievement of interdependent epistemic communities rather than isolated subjects, and progress is conceived as iterative and practice-dependent (Elgin 2025). However, developing such a framework in the context of AI—where the relevant community includes human and artificial contributors in structured sociotechnical arrangements—remains an open task.

completed at the moment one ticks the relevant boxes for fairness and trustworthiness (Funk 2024b). Our epistemological findings lend this concern a specific edge: if the epistemic standing of AI systems depends constitutively on the socio-technical practices surrounding them (external anchoring, scientific validation, institutional accountability), then regulatory frameworks that focus exclusively on the technical properties of the system itself—its risk classification, its transparency score, its bias metrics—will systematically miss the conditions under which AI outputs can function as genuine knowledge.

> How should AI governance frameworks be reconceived so as to regulate not merely the technical artifact but the epistemic practices—training, validation, deployment, correction—in which its outputs acquire (or fail to acquire) epistemic warrant?

(3) Our analysis also bears on the increasingly prominent debate about whether AI systems can be moral agents. Funk has identified what he terms the "posthumanist fallacy" (*posthumanistischer Fehlschluss*): the inferential error of placing technical information processing on a normative par with human moral agency, thereby inadvertently elevating a rationalist-reductionist worldview—formal language, propositional knowledge, means-end rationality implemented in technical artifacts—while simultaneously claiming to overcome precisely the "anthropocentric humanism" that produced it (ibid.). If our epistemological assessment is correct that CLERE's epistemic standing falls short of the conditions required for genuine knowledge—conditions that include answerability, normative participation, and embeddedness in a living practice—then the attribution of full moral agency to such systems appears even more difficult to sustain, since moral agency arguably presupposes at least the epistemic capacities our analysis has found wanting.

> To what extent does the epistemological deficit identified here—the absence of normative participation and answerability—constrain or preclude the attribution of moral agency to artificial neural networks, and what forms of distributed or relational moral responsibility are adequate to the socio-technical character of AI-mediated action?

(4) The finding that epistemic competence resides not in the isolated system but in the broader socio-technical practice raises questions that extend into the political philosophy of AI. Coeckelbergh has argued that the political dimensions of artificial intelligence cannot be adequately addressed by focusing on the properties of individual systems alone but require attention to the power relations, institutional structures, and governance mechanisms through which AI is embedded in social life (Coeckelbergh 2022). Our epistemological analysis supports and sharpens this claim: if what determines whether CLERE's outputs count as knowledge is not an intrinsic property of the system but rather the quality of the epistemic practices surrounding it—who selects the training data, who validates the outputs, who bears the consequences of

error—then the question of machine knowledge is inseparable from questions of epistemic power, authority, and justice.

> How should epistemic authority be distributed between human agents and AI systems within institutional decision-making, and what political and institutional safeguards are needed to ensure that the socio-technical practices on which AI's epistemic warrant depends remain transparent, contestable, and democratically accountable?

(5) Finally, our analysis brings into view a further question oriented toward future epistemic possibilities. If CLERE's knowledge is constitutively shaped by its computational form—as we argued, following Abel's insight that the forms through which knowledge is expressed are constitutive of, not merely instrumental for, its articulation (Abel and Conant 2012, pp. 6–7)—then the high-dimensional representational spaces in which ANNs operate may not merely approximate human epistemic achievements but may enable genuinely novel epistemic configurations. Peschl has investigated the conditions under which radically new knowledge can emerge, emphasizing that transformative innovation requires not merely recombination of existing elements but the creation of "enabling spaces" that allow fundamentally new patterns of understanding to take shape (Peschl 2012). Our findings suggest that the interaction between human epistemic agents and AI systems whose internal representations are organized along dimensions that no human could explicitly survey may constitute precisely such an enabling space—provided that the interaction is structured so as to preserve the conditions of epistemic responsibility identified in our analysis.

> Under what conditions can the collaboration between human knowers and AI systems—whose representational forms are irreducible to human-tractable propositions—give rise to genuinely novel knowledge that neither partner could have achieved alone, and what epistemological frameworks are needed to assess and govern such hybrid epistemic practices?

References

Abel, Günter and James Conant (2012). *Rethinking epistemology*. eng. Berlin studies in knowledge research 1. Berlin Boston: W. de Gruyter. ISBN: 978-3-11-025356-6.

Aggarwal, Charu C., Alexander Hinneburg, and Daniel A. Keim (2001). "On the Surprising Behavior of Distance Metrics in High Dimensional Space". In: *Database Theory — ICDT 2001*. Ed. by Jan Van den Bussche and Victor Vianu. Vol. 1973. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 420–434. DOI: 10.1007/3-540-44503-X_27.

Aristotle (2002). *Nicomachean Ethics: Translation, Introduction, Commentary*. Ed. by Sarah Broadie and Christopher Rowe. Trans. by Christopher Rowe. Oxford: Oxford University Press. ISBN: 9780198752714.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). "Neural Machine Translation by Jointly Learning to Align and Translate". In: *3rd International Conference on Learning Representations (ICLR) 2015, Conference Track Proceedings*. URL: http://arxiv.org/abs/1409.0473.

Bar-Hillel, Yehoshua (1960). "The Present Status of Automatic Translation of Languages". In: *Advances in Computers*. Ed. by Franz L. Alt and Morris Rubinoff. Vol. 1. Academic Press, pp. 91–163. DOI: 10.1016/S0065-2458(08)60607-5.

Barez, Fazl et al. (July 2025). *Chain-of-Thought Is Not Explainability*. Tech. rep. Preprint, under review. Oxford Martin AI Governance Initiative (University of Oxford). URL: https://aigi.ox.ac.uk/wp-content/uploads/2025/07/Cot_Is_Not_Explainability-1.pdf.

Barlow, Horace B. (1961). "Possible Principles Underlying the Transformations of Sensory Messages". In: *Sensory Communication*. Ed. by Walter A. Rosenblith. Cambridge, MA: MIT Press, pp. 217–234.

Bengio, Yoshua, Patrice Simard, and Paolo Frasconi (1994). "Learning Long-Term Dependencies with Gradient Descent is Difficult". In: *IEEE Transactions on Neural Networks* 5.2, pp. 157–166. DOI: 10.1109/72.279181.

Bergmann, Michael (2006). *Justification Without Awareness: A Defense of Epistemic Externalism*. New York: Oxford University Press.

Bergstrom, Carl T. and Jevin D. West (2026). *Case Study: Criminal Machine Learning*. Web case study (no publication date stated on the page). URL: https://callingbullshit.org/case_studies/case_study_criminal_machine_learning.html (visited on 01/23/2026).

Beyer, Kevin et al. (1999). "When Is "Nearest Neighbor" Meaningful?" In: *Database Theory — ICDT'99*. Ed. by Catriel Beeri and Peter Buneman. Vol. 1540. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, pp. 217–235. DOI: 10.1007/3-540-49257-7_15.

Block, Ned (1981). "Psychologism and Behaviourism". In: *The Philosophical Review* 90.1, pp. 5–43. DOI: 10.2307/2184371.

BonJour, Laurence (1980). "Externalist Theories of Empirical Knowledge". In: *Midwest Studies in Philosophy*. Ed. by Peter A. French, Theodore E. Uehling, and Howard K. Wettstein. Vol. 5. University of Minnesota Press, pp. 53–73. DOI: 10.1111/j.1475-4975.1980.tb00396.x.

Bostrom, Nick (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press. ISBN: 978-0-19-873983-8.

BruceBlaus (Sept. 30, 2013). *Blausen 0657 MultipolarNeuron*. Image. Source: own work. License: Creative Commons Attribution 3.0 Unported (CC BY 3.0). URL: https://commons.wikimedia.org/wiki/File:Blausen_0657_MultipolarNeuron.png (visited on 01/12/2026).

Burrell, Jenna (2016). "How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms". In: *Big Data & Society* 3.1, pp. 1–12. DOI: 10.1177/2053951715622512. URL: https://journals.sagepub.com/doi/10.1177/2053951715622512.

Cho, Kyunghyun et al. (2014). "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics. URL: https://aclanthology.org/D14-1179/.

Chollet, François (2019). *On the Measure of Intelligence*. Version Number: 2. DOI: 10.48550/ARXIV.1911.01547. URL: https://arxiv.org/abs/1911.01547 (visited on 08/08/2025).

Code, Lorraine (2006). *Ecological thinking: the politics of epistemic location*. eng. Studies in feminist philosophy. Oxford: Oxford University Press. ISBN: 978-0-19-978641-1. DOI: 10.1093/0195159438.001.0001.

Coeckelbergh, Mark (2022). *The Political Philosophy of AI: An Introduction*. Cambridge: Polity Press. ISBN: 978-1-509-54880-1.

Collins, Harry (2013). *Tacit and explicit knowledge*. eng. Chicago, Ill: University of Chicago Press. ISBN: 978-0-226-00421-1 978-0-226-11382-1.

Conee, Earl and Richard Feldman (1998). "The Generality Problem for Reliabilism". In: *Philosophical Studies* 89.1, pp. 1–29. DOI: 10.1023/A:1004243308503.

Cook, Roy (2024). "Frege's Logic". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Summer 2024. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/sum2024/entries/frege-logic/ (visited on 01/12/2026).

Coreth, Emerich and Harald Schöndorf (1983). *Philosophie des 17. und des 18. Jahrhunderts*. Stuttgart: W. Kohlhammer. ISBN: 3-17-008030-X.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-Vector Networks". In: *Neural Computation* 7.3, pp. 273–297. DOI: 10.1162/neco.1995.7.3.273.

D. O. Hebb (1949). *The Organization Of Behavior A Neuropsychological Theory*. eng. URL: http://archive.org/details/in.ernet.dli.2015.168156 (visited on 01/15/2026).

Descartes, René (1985). "Discourse on the Method of Rightly Conducting One's Reason and Seeking Truth in the Sciences". In: *The Philosophical Writings of Descartes*. Ed. and trans. by John Cottingham, Robert Stoothoff, and Dugald Murdoch. Vol. 1. Cambridge: Cambridge University Press, pp. 111–151. ISBN: 978-0-521-28807-1.

Dretske, Fred I. (Dec. 1970). "Epistemic Operators". en. In: *The Journal of Philosophy* 67.24, p. 1007. ISSN: 0022362X. DOI: 10.2307/2024710. URL: http://www.pdcnet.org/oom/service?url_ver=Z39.88-2004&rft_val_fmt=&rft.imuse_id=jphil_1970_0067_0024_1007_1023&svc_id=info:www.pdcnet.org/collection (visited on 01/10/2026).

Dreyfus, Hubert L. and Stuart E. Dreyfus (Jan. 1992). "What artificial experts can and cannot do". en. In: *AI & Soc* 6.1, pp. 18–26. ISSN: 0951-5666, 1435-5655. DOI: 10.1007/BF02472766. URL: http://link.springer.com/10.1007/BF02472766 (visited on 12/29/2025).

Durán, Juan Manuel (2025). *Beyond transparency: computational reliabilism as an externalist epistemology of algorithms*. Version Number: 1. DOI: 10.48550/ARXIV.2502.20402. URL: https://arxiv.org/abs/2502.20402 (visited on 11/24/2025).

Dusek, Val (2006). *Philosophy of Technology: An Introduction*. Malden, MA: Blackwell Publishing. ISBN: 978-1405111638.

Elgin, Catherine Z. (2017). *True Enough*. Cambridge, MA: MIT Press. ISBN: 978-0-262-03599-2.

– (2025). *Epistemic ecology*. eng. Cambridge (Mass.): MIT Press. ISBN: 978-0-262-55171-7.

Fawcett, Tom (June 2006). "An introduction to ROC analysis". en. In: *Pattern Recognition Letters* 27.8, pp. 861–874. ISSN: 01678655. DOI: 10.1016/j.patrec.2005.10.010. URL: https://linkinghub.elsevier.com/retrieve/pii/S016786550500303X (visited on 01/23/2026).

Feldman, Richard and Earl Conee (2001). “Internalism Defended”. en. In: *American Philosophical Quarterly* 38.1, pp. 1–18. URL: http://www.jstor.org/stable/20010019.

Firth, J. R. (1957). “A Synopsis of Linguistic Theory, 1930–1955”. In: *Studies in Linguistic Analysis*. Oxford: Basil Blackwell, pp. 1–32.

Funk, Michael (2022). *Roboter- und KI-Ethik: Eine methodische Einführung – Grundlagen der Technikethik Band 1*. de. Wiesbaden: Springer Fachmedien. ISBN: 978-3-658-34665-2 978-3-658-34666-9. DOI: 10.1007/978-3-658-34666-9. URL: https://link.springer.com/10.1007/978-3-658-34666-9 (visited on 02/01/2026).

– (2023). *Künstliche Intelligenz, Verkörperung und Autonomie: Theoretische Probleme - Grundlagen der Technikethik Band 4*. ger. Wiesbaden: Springer Vieweg. in Springer Fachmedien Wiesbaden GmbH. ISBN: 978-3-658-41105-3 978-3-658-41106-0.

– (2024a). “Nichtwissen ist Macht! Künstliche Intelligenz und ihre diversen Ethiken”. In: *Wissensdiversität und formatierte Bildungsräume / Medien – Wissen – Bildung*. Ed. by Andreas Beinsteiner, Ann-Kathrin Dittrich, and Theo Hug. innsbruck university press. ISBN: 978-3-99106-129-8. DOI: 10.15203/99106-129-8-04. URL: https://www.uibk.ac.at/iup/buch_pdfs/wissensdiversitaet-und-formatierte-bildungsraeume/10.15203-99106-129-8-04.pdf (visited on 02/15/2026).

– (2024b). “Zwischen AI Act und Posthumanismus. Künstliche Intelligenz und ihre ethischen „Radialkräfte“”. In: *Künstliche Intelligenz im Diskurs: Interdisziplinäre Perspektiven zur Gegenwart und Zukunft von KI-Anwendungen*. Ed. by Theo Hug, Petra Missomelius, and Heike Ortner. innsbruck university press. ISBN: 978-3-99106-139-7. DOI: 10.15203/99106-139-7-11. URL: https://www.uibk.ac.at/iup/buch_pdfs/ki-im-diskurs/10.15203-99106-139-7-11.pdf (visited on 02/15/2026).

Gardner, Sebastian (1999). *Routledge Philosophy GuideBook to Kant and the Critique of Pure Reason*. London: Routledge.

Gascoigne, Neil and Timothy Thornton (2013). *Tacit knowledge*. eng. Durham: Acumen. ISBN: 978-1-84465-546-5.

Geirhos, Robert et al. (Nov. 2020). “Shortcut learning in deep neural networks”. en. In: *Nat Mach Intell* 2.11, pp. 665–673. ISSN: 2522-5839. DOI: 10.1038/s42256-020-00257-z. URL: https://www.nature.com/articles/s42256-020-00257-z (visited on 07/30/2025).

Gerken, Mikkel (2017). *On folk epistemology: how we think and talk about knowledge*. eng. Oxford: Oxford University Press. ISBN: 978-0-19-880345-4.

Gitelman, Lisa, ed. (2013). *“Raw Data” Is an Oxymoron*. Cambridge, MA: The MIT Press.

Goldman, Alvin and Bob Beddor (2021). "Reliabilist Epistemology". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2021. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/sum2021/entries/reliabilism/ (visited on 12/17/2025).

Goldman, Alvin I. (1999). *Knowledge in a Social World*. Oxford: Oxford University Press. ISBN: 978-0-19-823820-4.

Goodfellow, Ian, Aaron Courville, and Yoshua Bengio (2016). *Deep learning*. eng. Adaptive computation and machine learning. Cambridge, Massachusetts: The MIT Press. ISBN: 978-0-262-03561-3 978-0-262-33737-3.

Gramelsberger, Gabriele and Hans-Jörg Sigwart (2025). *Epistemologien zur Einführung*. ger. Zur Einführung. Hamburg: Junius. ISBN: 978-3-96060-350-4.

Grimm, Stephen R. (2011). "Understanding". In: ed. by Sven Bernecker and Duncan Pritchard, pp. 84–94.

– (2021). *Understanding*. Internet Encyclopedia of Philosophy. URL: https://iep.utm.edu/understa/ (visited on 02/19/2026).

Harris, Keith Raymond (Sept. 2024). "Ability, Knowledge, and Non-paradigmatic Testimony". en. In: *Episteme* 21.3, pp. 983–1001. ISSN: 1742-3600, 1750-0117. DOI: 10.1017/epi.2023.2. URL: https://www.cambridge.org/core/product/identifier/S1742360023000023/type/journal_article (visited on 01/10/2026).

Harris, Zellig S. (1954). "Distributional Structure". In: *Word* 10.2–3, pp. 146–162. DOI: 10.1080/00437956.1954.11659520.

Haugeland, John (1993). *Artificial intelligence: the very idea*. 6. print. Bradford books. Cambridge, Mass.: MIT Press. ISBN: 978-0-262-08153-5 978-0-262-58095-3.

Héder, Mihály, Daniel Paksi, and The Polanyi Society (2018). "Non-Human Knowledge According to Michael Polanyi:" en. In: *Tradition and Discovery: The Polanyi Society Periodical* 44.1, pp. 50–66. ISSN: 1057-1027. DOI: 10.5840/traddisc20184418. URL: http://www.pdcnet.org/oom/service?url_ver=Z39.88-2004&rft_val_fmt=&rft.imuse_id=traddisc_2018_0044_0001_0050_0066&svc_id=info:www.pdcnet.org/collection (visited on 01/23/2026).

Herczeg, Petra et al. (2023). *Guidelines der Universität Wien zum Umgang mit Künstlicher Intelligenz (KI) in der Lehre: 2. Auflage 2024*. de. Artwork Size: 3219192 b Medium: application/pdf. DOI: 10.25365/PHAIDRA.544. URL: https://phaidra.univie.ac.at/o:2092606 (visited on 02/19/2026).

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

Hofmann, Thomas (Dec. 2022). "Deep Learning". Publisher: ETH Zurich, Department of Computer Science.

Hopcroft, John E., Rajeev Motwani, and Jeffrey D. Ullman (2006). *Introduction to Automata Theory, Languages, and Computation*. 3rd ed. Boston: Pearson. ISBN: 9780321455369.

Hornik, Kurt, Maxwell Stinchcombe, and Halbert White (Jan. 1989). "Multilayer feedforward networks are universal approximators". en. In: *Neural Networks* 2.5, pp. 359–366. ISSN: 08936080. DOI: 10.1016/0893-6080(89)90020-8. URL: https://linkinghub.elsevier.com/retrieve/pii/0893608089900208 (visited on 01/15/2026).

Hume, David (1978). *A Treatise of Human Nature*. Ed. by L. A. Selby-Bigge. 2nd ed. 2nd ed., revised with variant readings by P. H. Nidditch. Oxford: Clarendon Press.

Ichikawa, Jonathan and Matthias Steup (2024). "The Analysis of Knowledge". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2024. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/fall2024/entries/knowledge-analysis/ (visited on 01/09/2026).

Jackson, Peter and Isabelle Moulinier (2007). *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*. 2nd ed. Vol. 5. Natural Language Processing. Amsterdam and Philadelphia: John Benjamins Publishing Company. ISBN: 9789027292445. DOI: 10.1075/nlp.5.

Jurafsky, Daniel and James H. Martin (2026). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd ed. Online manuscript, draft of January 6, 2026. URL: https://web.stanford.edu/~jurafsky/slp3/.

Kanai, Ryota, Ryota Takatsuki, and Ippei Fujisawa (2025). "Meta-representations as representations of processes". In: *Neuroscience of Consciousness* 2025.1, niaf038. DOI: 10.1093/nc/niaf038. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC12573243/.

Kandel, Eric R. et al. (2021). *Principles of Neural Science*. 6th ed. McGraw Hill. ISBN: 9781259642234.

Kant, Immanuel (Jan. 2018). *Critique of Pure Reason*. eng. [S.l.]: Seltzer Books. ISBN: 978-1-4553-2297-8. URL: https://research.ebsco.com/linkprocessor/plink?id=41ca1d21-64cc-34b3-b5de-d4d6913b4835 (visited on 01/10/2026).

Kelp, Christoph (2015). "Understanding Phenomena". In: *Synthese* 192.12, pp. 3799–3816. DOI: 10.1007/s11229-014-0616-x.

Khandelwal, Urvashi et al. (2020). "Generalization through Memorization: Nearest Neighbor Language Models". In: *International Conference on Learning Representations (ICLR)*. arXiv: 1911.00172 [cs.CL]. URL: https://openreview.net/forum?id=HklBjCEKvH.

Kim, Jaegwon (1996). *Philosophy of mind*. eng. Dimensions of philosophy series. Boulder (Colo.) Oxford (GB): Westview press. ISBN: 978-0-8133-0775-6.

Kirkham, Richard L. (1992). *Theories of Truth: A Critical Introduction*. A Bradford Book. Cambridge, MA: MIT Press. ISBN: 978-0-262-11167-6.

Kirkpatrick, James et al. (2017). "Overcoming catastrophic forgetting in neural networks". In: *Proceedings of the National Academy of Sciences of the United States of America* 114.13, pp. 3521–3526. DOI: 10.1073/pnas.1611835114. URL: https://pubmed.ncbi.nlm.nih.gov/28292907/.

Kitchin, Rob and Tracey P. Lauriault (2014). *Towards Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work*. Working Paper 2. The Programmable City (Maynooth University). URL: https://ssrn.com/abstract=2474112 (visited on 01/15/2026).

Korcz, Keith Allen (2021). "The Epistemic Basing Relation". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2021. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/spr2021/entries/basing-epistemic/ (visited on 01/10/2026).

Kulstad, Mark and Laurence Carlin (2020). "Leibniz's Philosophy of Mind". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Winter 2020. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/win2020/entries/leibniz-mind/.

Lagerlund, Henrik (2022). "Medieval Theories of the Syllogism". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Summer 2022. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/sum2022/entries/medieval-syllogism/ (visited on 01/12/2026).

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *Nature* 521, pp. 436–444. DOI: 10.1038/nature14539.

Legg, Shane and Marcus Hutter (2007). "A Collection of Definitions of Intelligence". In: *Frontiers in Artificial Intelligence and Applications* 157, pp. 17–24.

Leibniz, Gottfried Wilhelm (1902). *Discourse on Metaphysics, Correspondence with Arnauld, and Monadology*. Trans. by George R. Montgomery. Chicago: The Open Court Publishing Company. URL: https://archive.org/download/cu31924014598829/cu31924014598829.pdf (visited on 01/09/2026).

Leibniz, Gottfried Wilhelm (1969). *Philosophical Papers and Letters*. Ed. and trans. by Leroy E. Loemker. 2nd ed. Vol. 2. Synthese Historical Library. Dordrecht: D. Reidel.

Lichtman, Jeff W. and Winfried Denk (Nov. 2011). "The Big and the Small: Challenges of Imaging the Brain's Circuits". en. In: *Science* 334.6056, pp. 618–623. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1209168. URL: https://www.science.org/doi/10.1126/science.1209168 (visited on 01/23/2026).

Lin, Hanti (2024). "Bayesian Epistemology". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Summer 2024. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/sum2024/entries/epistemology-bayesian/ (visited on 01/10/2026).

Littlejohn, Clayton (2025). "Internalist vs. Externalist Conceptions of Epistemic Justification". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Winter 2025. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/win2025/entries/justep-intext/ (visited on 01/10/2026).

Longino, Helen E. (2002). *The fate of knowledge*. eng. Princeton, N.J: Princeton University Press. ISBN: 978-0-691-08876-1 978-0-691-08875-4 978-0-691-18701-3.

Lyons, Jack C. (2009). *Perception and Basic Beliefs: Zombies, Modules, and the Problem of the External World*. New York: Oxford University Press. DOI: 10.1093/acprof:oso/9780195373578.001.0001.

Manning, Christopher D. and Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz (1993). "Building a Large Annotated Corpus of English: The Penn Treebank". In: *Computational Linguistics* 19.2, pp. 313–330.

McCulloch, Warren S. and Walter Pitts (Dec. 1943). "A Logical Calculus of the Ideas Immanent in Nervous Activity". In: *The Bulletin of Mathematical Biophysics* 5.4, pp. 115–133. DOI: 10.1007/BF02478259. URL: https://home.csulb.edu/~cwallis/382/readings/482/mccolloch.logical.calculus.ideas.1943.pdf (visited on 01/12/2026).

Melamed, Yitzhak Y. and Martin Lin (2023). "Principle of Sufficient Reason". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Summer 2023. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/sum2023/entries/sufficient-reason/.

Merriam-Webster (n.d.). *Datum*. Dictionary entry. Merriam-Webster. URL: https://www.merriam-webster.com/dictionary/datum (visited on 01/15/2026).

Mikolov, Tomas et al. (2013). "Efficient Estimation of Word Representations in Vector Space". In: *arXiv preprint arXiv:1301.3781*. arXiv: 1301.3781 [cs.CL]. URL: https://arxiv.org/abs/1301.3781.

Millière, Raphaël and Cameron Buckner (2024). *A Philosophical Introduction to Language Models – Part I: Continuity With Classic Debates*. Version Number: 1. DOI: 10.48550/ARXIV.2401.03910. URL: https://arxiv.org/abs/2401.03910 (visited on 05/29/2025).

Mitchell, Tom M. (1997). *Machine Learning*. 1st ed. New York: McGraw-Hill Education. ISBN: 978-0-07-042807-2.

Mugleston, Jennifer et al. (Feb. 2025). "Epistemology in the Age of Large Language Models". en. In: *Knowledge* 5.1, p. 3. ISSN: 2673-9585. DOI: 10.3390/knowledge5010003. URL: https://www.mdpi.com/2673-9585/5/1/3 (visited on 05/29/2025).

Murphy, Kevin P. (2012). *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.

Nagel, Sebastian and Thom Vaughan (Dec. 18, 2024). *December 2024 Crawl Archive Now Available*. Common Crawl. URL: https://commoncrawl.org/blog/december-2024-crawl-archive-now-available (visited on 01/15/2026).

National Research Council (U.S.). Automatic Language Processing Advisory Committee (1966). *Language and Machines: Computers in Translation and Linguistics*. Washington, DC: The National Academies Press. DOI: 10.17226/9547.

Newell, Allen and Herbert A. Simon (Sept. 1956). "The Logic Theory Machine—A Complex Information Processing System". In: *IRE Transactions on Information Theory* 2.3, pp. 61–79. DOI: 10.1109/TIT.1956.1056797. URL: https://ieeexplore.ieee.org/document/1056797 (visited on 01/12/2026).

– (Mar. 1976). "Computer science as empirical inquiry: symbols and search". en. In: *Commun. ACM* 19.3, pp. 113–126. ISSN: 0001-0782, 1557-7317. DOI: 10.1145/360018.360022. URL: https://dl.acm.org/doi/10.1145/360018.360022 (visited on 01/12/2026).

– (1988). "GPS, A Program That Simulates Human Thought". en. In: *Readings in Cognitive Science*. Elsevier, pp. 453–460. ISBN: 978-1-4832-1446-7. DOI: 10.1016/B978-1-4832-1446-7.50040-6. URL: https://linkinghub.elsevier.com/retrieve/pii/B9781483214467500406 (visited on 09/18/2025).

O'Connor, Cailin, Sanford Goldberg, and Alvin Goldman (2024). "Social Epistemology". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Summer 2024. Metaphysics Research Lab, Stanford University. URL: https://plato.stanford.edu/archives/sum2024/entries/epistemology-social/.

Otsuka, Jun (Nov. 2022). *Thinking About Statistics: The Philosophical Foundations*. en. 1st ed. New York: Routledge. ISBN: 978-1-003-31906-1. DOI: 10.4324/9781003319061. URL: https://www.taylorfrancis.com/books/9781003319061 (visited on 10/15/2025).

Parisi, German I. et al. (May 2019). "Continual lifelong learning with neural networks: A review". In: *Neural Networks* 113, pp. 54–71. DOI: 10.1016/j.neunet.2019.01.012. URL: https://www.sciencedirect.com/science/article/pii/S0893608019300231.

Pascanu, Razvan, Tomas Mikolov, and Yoshua Bengio (2013). "On the difficulty of training recurrent neural networks". In: *Proceedings of the 30th International Conference on Machine Learning (ICML)*. Often cited for vanishing/exploding gradients in RNNs; arXiv:1211.5063. URL: https://arxiv.org/abs/1211.5063.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543. URL: https://nlp.stanford.edu/pubs/glove.pdf.

Peschl, Markus (2012). "Spaces enabling game-changing and sustaining innovations: why space matters for knowledge creation and innovation". In: *jotsc* 9.1. ISSN: 14779633. DOI: 10.1386/otsc.9.1.41_1. URL: http://www.ingentaconnect.com/content/intellect/jotsc/2012/00000009/00000001/art00004 (visited on 02/15/2026).

Peters, Matthew E. et al. (2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 2227–2237. URL: https://aclanthology.org/N18-1202/.

Polanyi, Michael (1967). *The Tacit Dimension*. en. Google-Books-ID: Hd6SxgEACAAJ. Anchor Books. ISBN: 978-0-14-341418-6.

– (1998). *Personal knowledge: towards a post-critical philosophy*. eng. London: Routledge. ISBN: 978-0-415-15149-8.

Poston, Ted (2026). *Internalism and Externalism in Epistemology*. Internet Encyclopedia of Philosophy (ISSN 2161-0002). The IEP does not provide publication or revision dates; cite by access date. Internet Encyclopedia of Philosophy. URL: https://iep.utm.edu/int-ext/ (visited on 01/10/2026).

Radford, Alec et al. (2018). *Improving Language Understanding by Generative Pre-Training*. Tech. rep. OpenAI technical report. URL: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Räuker, Tilman et al. (2022). *Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks*. Version Number: 6. DOI: 10.48550/ARXIV.2207.13243. URL: https://arxiv.org/abs/2207.13243 (visited on 07/06/2025).

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

Rodriguez-Pereyra, Gonzalo (2000). "What is the Problem of Universals?" In: *Mind* 109.434. Publisher: [Oxford University Press, Mind Association], pp. 255–273. ISSN: 0026-4423. URL: https://www.jstor.org/stable/2660134 (visited on 01/09/2026).

Rolnick, David et al. (2019). "Experience Replay for Continual Learning". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS 2019), December 8–14, 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach et al., pp. 348–358. URL: https://proceedings.neurips.cc/paper/2019/hash/fa7cdfad1a5aaf8370ebeda47a1ff1c3-Abstract.html.

Rosenblatt, F. (1958). "The perceptron: A probabilistic model for information storage and organization in the brain." en. In: *Psychological Review* 65.6, pp. 386–408. ISSN: 1939-1471, 0033-295X. DOI: 10.1037/h0042519. URL: https://doi.apa.org/doi/10.1037/h0042519 (visited on 01/15/2026).

Rosenblatt, Frank (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. See ch. 5, "The Principal Convergence Theorem" (Rosenblatt's convergence result for the perceptron). Washington, DC: Spartan Books. URL: https://gwern.net/doc/ai/nn/1962-rosenblatt-principlesofneurodynamics.pdf.

Russell, Bertrand (1912). *The Problems of Philosophy*. Home University Library of Modern Knowledge 35. Internet Archive scan of the 1912 edition. New York and London: Henry Holt et al. URL: https://ia902804.us.archive.org/5/items/problemsofphilo00russuoft/problemsofphilo00russuoft.pdf (visited on 01/10/2026).

Russell, Stuart and Peter Norvig (2016). *Artificial Intelligence: A Modern Approach*. ger. 3. Aufl. Erscheinungsort nicht ermittelbar. ISBN: 978-1-292-15396-4.

Rusu, Andrei A. et al. (2016). "Progressive Neural Networks". In: *arXiv preprint*. DOI: 10.48550/arXiv.1606.04671. arXiv: 1606.04671 [cs.LG]. URL: https://arxiv.org/abs/1606.04671.

Ryle, Gilbert (1946). "Knowing How and Knowing That: The Presidential Address". In: *Proceedings of the Aristotelian Society*. New Series 46, pp. 1–16. DOI: 10.1093/aristotelian/46.1.1. URL: http://www.jstor.org/stable/4544405.

– (2009). *The Concept of Mind*. 60th Anniversary Edition. Originally published 1949; this edition includes an introduction by Julia Tanney. London and New York: Routledge. ISBN: 9780415485470.

Sartwell, Crispin (1991). "Knowledge Is Merely True Belief". In: *American Philosophical Quarterly* 28.2. Publisher: [North American Philosophical Publications, University of Illinois Press], pp. 157–165. ISSN: 0003-0481. URL: https://www.jstor.org/stable/20014367 (visited on 01/09/2026).

Schölkopf, Bernhard and Alexander J. Smola (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press. ISBN: 9780262194754.

Searle, John R. (Sept. 1980). "Minds, brains, and programs". en. In: *Behav Brain Sci* 3.3, pp. 417–424. ISSN: 0140-525X, 1469-1825. DOI: 10.1017/S0140525X00005756. URL: https://www.cambridge.org/core/product/identifier/S0140525X00005756/type/journal_article (visited on 01/09/2026).

Sellars, Wilfrid (1956). "Empiricism and the Philosophy of Mind". In: *Minnesota Studies in the Philosophy of Science, Volume I: The Foundations of Science and the Concepts of Psychology and Psychoanalysis*. Ed. by Herbert Feigl and Michael Scriven. Minneapolis: University of Minnesota Press, pp. 253–329.

Shalev-Shwartz, Shai and Shai Ben-David (May 2014). *Understanding Machine Learning: From Theory to Algorithms*. en. 1st ed. Cambridge University Press. ISBN: 978-1-107-05713-5 978-1-107-29801-9. DOI: 10.1017/CBO9781107298019. URL: https://www.cambridge.org/core/product/identifier/9781107298019/type/book (visited on 01/15/2026).

Shannon, Claude E. (1938). "A Symbolic Analysis of Relay and Switching Circuits". In: *Transactions of the American Institute of Electrical Engineers* 57. Paper states it is an abstract of an MIT M.S. thesis, pp. 713–723. DOI: 10.1109/T-AIEE.1938.5057767.

– (1948). "A Mathematical Theory of Communication". In: *The Bell System Technical Journal* 27.3. Reprint PDF, pp. 379–423. DOI: 10.1002/j.1538-7305.1948.tb01338.x. URL: https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf (visited on 01/15/2026).

Shapiro, Lawrence (2019). *Embodied Cognition*. 2nd ed. Routledge. ISBN: 9781138746992. DOI: 10.4324/9781315180380.

Shatz, Carla J. (Sept. 1992). "The Developing Brain". In: *Scientific American* 267.3, pp. 60–67. DOI: 10.1038/scientificamerican0992-60.

Silva, Paul and Luis R.G. Oliveira (Nov. 2023). "Propositional Justification and Doxastic Justification". en. In: *The Routledge Handbook of the Philosophy of Evidence*. Ed. by Maria Lasonen-Aarnio and Clayton Littlejohn. 1st ed. London: Routledge, pp. 395–407. ISBN: 978-1-315-67268-7. DOI: 10.4324/9781315672687-36. URL: https://www.taylorfrancis.com/books/9781315672687/chapters/10.4324/9781315672687-36 (visited on 01/10/2026).

Simon, Herbert Alexander (2008). *The sciences of the artificial*. en. 3. ed., [Nachdr.] Cambridge, Mass.: MIT Press. ISBN: 978-0-262-19374-0 978-0-262-69191-8.

Simoncelli, Eero P. and Bruno A. Olshausen (2001). "Natural Image Statistics and Neural Representation". In: *Annual Review of Neuroscience* 24, pp. 1193–1216. DOI: 10.1146/annurev.neuro.24.1.1193.

Starmer, Joshua (2025). *The StatQuest Illustrated Guide to Neural Networks and AI: A Visual, Interactive Guide to Machine Learning, Deep Learning and Generative AI*. Packt Publishing. ISBN: 9781835084967.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le (2014). "Sequence to Sequence Learning with Neural Networks". In: *Advances in Neural Information Processing Systems*. NeurIPS 2014. URL: https://www.semanticscholar.org/paper/cea967b59209c6be22829699f05b8b1ac4dc092d.

Swift, Jonathan (2005). *Gulliver's Travels*. Ed. by Claude Rawson and Ian Higgins. Originally published 1726. Oxford: Oxford University Press. ISBN: 978-0-19-280534-8.

Tanney, Julia (2012). "Ryle's Regress and the Philosophy of Cognitive Science". In: *Rules, Reason, and Self-Knowledge*. Cambridge, MA: Harvard University Press, pp. 249–276. DOI: 10.4159/harvard.9780674067837.

Tishby, Naftali, Fernando C. Pereira, and William Bialek (Apr. 2000). *The information bottleneck method*. arXiv:physics/0004057. DOI: 10.48550/arXiv.physics/0004057. URL: http://arxiv.org/abs/physics/0004057 (visited on 01/16/2026).

Torralba, Antonio, Phillip Isola, and William T. Freeman (2026). *Backpropagation*. See Fig. 14.3 "Basic sequential computation graph". URL: https://visionbook.mit.edu/backpropagation.html (visited on 01/15/2026).

Tribus, Myron and Edward C. McElroy (1971). "Energy and Information". In: *Scientific American* 224.3, pp. 179–188.

Turing, Alan M. (1950). "Computing Machinery and Intelligence". In: *Mind* 59.236, pp. 433–460. ISSN: 0026-4423. DOI: 10.1093/mind/LIX.236.433.

Turpin, Miles et al. (2023). "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting". In: *Advances in Neural Information Processing Systems*. arXiv: 2305.04388 [cs.CL]. URL: https://arxiv.org/abs/2305.04388.

U.S. Government Accountability Office (Mar. 2009). *Nuclear Weapons: NNSA and DOD Need to More Effectively Manage the Stockpile Life Extension Program*. GAO-09-385. Washington, DC: U.S. Government Accountability Office. URL: https://www.gao.gov/assets/gao-09-385.pdf (visited on 01/11/2026).

Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pp. 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Virani, Nurali, Naresh Iyer, and Zhaoyuan Yang (2019). *Justification-Based Reliability in Machine Learning*. Version Number: 3. DOI: 10.48550/ARXIV.1911.07391. URL: https://arxiv.org/abs/1911.07391 (visited on 11/19/2025).

Voita, Elena (Sept. 2020). *NLP Course For You*. URL: https://lena-voita.github.io/nlp_course.html.

Wagemans, Johan et al. (2012). "A Century of Gestalt Psychology in Visual Perception: I. Perceptual Grouping and Figure-Ground Organization". In: *Psychological Bulletin* 138.6, pp. 1172–1217. DOI: 10.1037/a0029333.

Wallace, David (2012). *The Emergent Multiverse: Quantum Theory according to the Everett Interpretation*. Oxford: Oxford University Press. ISBN: 978-0-19-954696-1.

Wasserman, Larry (2004). *All of Statistics. A Concise Course in Statistical Inference*. Springer Texts in Statistics. New York: Springer. ISBN: 978-0-387-40272-7. DOI: 10.1007/978-0-387-21736-9.

Wittgenstein, Ludwig (1922). *Tractatus Logico-Philosophicus*. Trans. by C. K. Ogden. With an intro. by Bertrand Russell. German text with English translation on facing pages. London: Routledge & Kegan Paul.

Wu, Xiaolin and Xi Zhang (2016). *Automated Inference on Criminality using Face Images*. arXiv preprint (v1), submitted 13 Nov 2016. DOI: 10.48550/arXiv.1611.04135. arXiv: 1611.04135 [cs.CV]. URL: https://arxiv.org/abs/1611.04135.

Zech, John R. et al. (Nov. 2018). "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study". en. In: *PLoS Med* 15.11. Ed. by Aziz Sheikh, e1002683. ISSN: 1549-1676. DOI: 10.1371/journal.pmed.1002683. URL: https://dx.plos.org/10.1371/journal.pmed.1002683 (visited on 01/07/2026).

Zipf, George Kingsley (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, MA: Addison-Wesley Press.

Zwiebach, Barton (2009). *A First Course in String Theory*. 2nd ed. Cambridge: Cambridge University Press. ISBN: 9780521880329.

List of Figures

A. Appendix

A.1. Declaration on the Use of AI Tools

The following table documents all AI-assisted tools used in the preparation of this thesis in accordance with the University of Vienna guidelines on the use of AI in academic work (Herczeg et al. 2023).

AI Tool	**Use**	**Parts of Work Affected**	**Remarks**
DeepL	Language assistance: translation, grammar correction, and linguistic formulation	Throughout the thesis	English is not the author's native language. DeepL was used to improve the linguistic quality of formulations. All intellectual content, arguments, and conclusions are entirely the author's own. The AI-generated suggestions were reviewed and revised independently.
QuillBot	Language assistance: grammar checking, proofreading, and linguistic formulation	Throughout the thesis	QuillBot was used as a grammar checker to improve clarity and correctness of English formulations. All intellectual content, arguments, and conclusions are entirely the author's own. The tool's suggestions were reviewed and revised independently.
Claude (Anthropic)	Assistance with the technical implementation of figures (LaTeX/Ti*k*Z code generation)	Figs. 2.1, 2.4, 2.6, 2.7, 2.8, 2.11, and 2.12	The figures were conceptually designed by the author. Claude was used as a coding aid for the LaTeX/Ti*k*Z implementation.
ChatGPT o3 (thinking) (OpenAI)	Assistance with the technical implementation of figures (LaTeX/Ti*k*Z code generation)	Figs. 2.1, 2.4, 2.6, 2.7, 2.8, 2.11, and 2.12	The figures were conceptually designed by the author. ChatGPT was used as a coding aid for the LaTeX/Ti*k*Z implementation.

Table A.1.: *Documentation of AI tool usage in accordance with Herczeg et al. (2023).*