Psychology Nexus

www.nex.reapress.com

Psych. Nex. Vol. 3, No. 1 (2026) 11–26.

Paper Type: Original Article

Belief, Bias, the Bayesian Machine: A Philosophical Probabilistic Inquiry into Human Cognitive Inference and Artificial Intelligence

Janardan Behera*

Department of Statistics, Ravenshaw University, Cuttack, Odisha, India; janardan@ravenshawuniversity.ac.in.

Citation:

Received: 10 September 2025 *Revised: 13 November 2025* *Accepted: 21 January 2026*	Behera, J. (2026). Belief, Bias, the Bayesian machine: A philosophical probabilistic inquiry into human cognitive inference and artificial intelligence. *Psychology Nexus, 3*(1), 11-26.

Abstract

Every mind, whether biological or artificial, is built not upon certainty but upon the art of living with uncertainty. This article develops a philosophically grounded and conceptually rigorous framework for understanding both human cognition and Artificial Intelligence (AI) as fundamentally probabilistic enterprises, systems perpetually engaged in estimating the plausibility of things, revising those estimates when evidence arrives, and acting under the permanent shadow of incomplete knowledge. Drawing on the Bayesian philosophy of mind, the heuristics-and-biases tradition, predictive coding neuroscience, and probabilistic machine learning, the paper argues that the most intellectually productive comparison between human thought and machine inference is not a competition of accuracy but a structural study in contrasting styles of uncertainty management. Human cognition is anchored in priors forged through embodied biography, cultural inscription, and emotion; it weighs evidence through the felt texture of personal history; and it generates hypotheses that exceed the statistical boundaries of any training corpus. AI, by contrast, operates as a formally honest bookkeeper of uncertainty, updating beliefs with mechanical consistency, carrying no affective debt, and retaining incapable of questioning the epistemic adequacy of its own foundational assumptions. The article introduces the original concept of the uncertainty signature as a six-dimensional philosophical diagnostic for characterizing how any reasoning system relates to the unknown. The analysis reveals that the deepest divergence between mind and machine is not computational but ontological: The human being is uncertain about itself, whereas the machine is uncertain only about the world. The philosophical, epistemic, and social implications of this divergence are examined in depth, including the risk of epistemic colonization when machine confidence systematically displaces human deliberative judgment and the conditions under which human and artificial reasoners can form genuinely complementary cognitive partnerships.

Keywords: Probabilistic cognition, Bayesian epistemology, Artificial intelligence, Uncertainty, Bounded rationality, Epistemic humility, Cognitive bias, Prior belief, Predictive coding, Human-machine complementarity.

1 | Introduction

The question of how a mind forms beliefs under uncertainty is among the oldest and most persistently unresolved problems in philosophy. It is also, increasingly, among the most practically consequential. As

Corresponding Author: janardan@ravenshawuniversity.ac.in

https://doi.org/10.48314/nex.v3i1.31

Artificial Intelligence (AI) systems assume decision-making roles in medicine, law, finance, and governance, the comparison between human cognitive inference and machine probabilistic reasoning has moved from an academic curiosity to an urgent intellectual and ethical necessity. Yet despite the proliferation of technical literature on machine learning and the parallel expansion of cognitive science research on human judgment, the two traditions have rarely engaged each other at the level of foundational philosophy. This article attempts precisely that engagement.

Probability theory, in its deepest philosophical reading, is not merely a branch of mathematics applied to games of chance or statistical models. It is a normative framework for rational existence under ignorance. The foundational arguments of Ramsey [1], De Finetti [2], and Savage [3] established that any agent holding coherent beliefs about an uncertain world must behave as if assigning numerical probabilities to outcomes. This result, sometimes called the Dutch book theorem, is not a prescription for behavior; it is a characterization of what internal coherence demands. Jaynes and Bretthorst [4] extended this foundation into an explicit epistemology, arguing that probability theory is the uniquely consistent extension of classical logic to situations of incomplete information. On this reading, every system that maintains beliefs and acts upon them, whether a human being reasoning about a diagnosis or a neural network classifying an image, is either implicitly probabilistic or implicitly incoherent.

The implications of this claim for cognitive science were developed systematically by Oaksford and Chater [5], who argued that human reasoning in natural contexts is better characterized as approximate probabilistic inference than as the application of deductive logical rules. The prior-to-posterior updating model that Tenenbaum et al. [6] demonstrated in developmental studies, and the predictive coding framework elaborated by Friston [7], together provide strong empirical support for the view that the brain implements something functionally equivalent to Bayesian inference, not as a deliberate computational strategy but as its basic operating architecture. The question, then, is not whether human cognition is probabilistic but what kind of probabilistic system it is and how it compares with the explicit probabilistic machinery of contemporary AI.

On the artificial side, the emergence of probabilistic machine learning as the dominant paradigm in AI research has produced a family of systems whose relationship to uncertainty is formally transparent. Ghahramani [8] provided a programmatic account of probabilistic machine learning as the theoretical foundation for AI, arguing that representing and reasoning about uncertainty is the central unresolved challenge of the field. The success of Bayesian deep learning, Gaussian processes, and variational inference methods documented in Bishop and Nasrabadi [9] and extended in the variational framework of Blei et al. [10] has produced systems that can express calibrated uncertainty over their predictions in ways that exceed the typical capacity of unaided human experts. At the same time, the well-documented failures of these systems under distributional shift, the brittleness exposed by Goodfellow et al. [11], and the calibration failures catalogued by Guo et al. [12] reveal that formal probabilistic architecture is neither sufficient for nor equivalent to the kind of robust, flexible, self-correcting cognition that human beings exhibit across ecologically rich and continuously changing environments.

The philosophical gap between these two traditions, the cognitive science of human probabilistic reasoning and the computer science of machine probabilistic inference, has been noted by several authors but rarely addressed at a conceptual depth adequate to the question's importance. Dreyfus [13] identified, decades before the current wave of machine learning, the fundamental inadequacy of formalized knowledge representations as a model of human intelligence, arguing that expertise is grounded in embodied engagement with the world rather than symbol manipulation. Clark [14] revisited this theme in the context of predictive processing, showing that the brain's probabilistic machinery is inseparable from its bodily and environmental situatedness. Russell [15] approached the comparison from the engineering side, arguing that the alignment of machine objectives with human values requires a far deeper understanding of how human preferences are probabilistically structured than current AI design acknowledges.

Against this background, the present article makes four original contributions. It develops a philosophically coherent account of both human cognition and AI as species of probabilistic inference that differ in their

prior structure, their approximation strategies, their phenomenological texture, and their capacity for reflexive self-revision. It introduces the concept of the Uncertainty Signature as a structured six-dimensional diagnostic for characterizing these differences in philosophically precise terms. It provides a systematic reinterpretation of cognitive biases and machine learning pathologies as parallel, structurally analogous disruptions of probabilistic rationality that share a common theoretical root. It examines the social and epistemic consequences of deploying machine probabilism in human institutions and articulates the conditions under which a philosophically serious human-AI cognitive partnership becomes possible.

The remainder of the paper is structured as follows. Section 2 reviews the relevant literature in four thematic areas: 1) The philosophy of probability, 2) The cognitive science of human probabilistic reasoning, 3) The probabilistic foundations of AI, and 4) The comparative philosophy of mind. Section 3 develops the philosophical model of human cognition as embodied Bayesian inference. Section 4 characterizes AI as a formally explicit but philosophically impoverished probabilistic reasoner. Section 5 introduces the Uncertainty Signature and applies it to both systems. Section 6 analyses cognitive biases and machine learning failures as structurally analogous phenomena. Section 7 examines the convergence of biological and artificial cognition through the lens of prediction, surprise, and free energy. Section 8 addresses the ethical and social dimensions of the comparison. Section 9 concludes with reflections on intelligence, wisdom, and the irreducibly uncertain condition of any mind.

2 | Literature Review

2.1 | The Philosophy of Probability and Rational Belief

The philosophical foundations of probability theory bear directly on the comparison between human and artificial reasoning because they determine what it means for any system to manage uncertainty rationally. The two principal traditions, frequentism and subjectivism, have profoundly different implications for how probabilistic cognition should be understood and evaluated.

The frequentist tradition, associated with Von Mises [16], holds that probability is a property of repeatable physical processes: The limiting frequency of an outcome in an infinite series of trials. On this account, probability statements about singular events, such as the probability that a particular patient has a particular diagnosis, are either meaningless or require elaborate reference-class arguments to make tractable. The frequentist framework dominated scientific statistics for much of the twentieth century and continues to inform standard null hypothesis significance testing. However, as Jaynes and Bretthorst [4] argued at length, the frequentist account cannot serve as a foundation for the kind of singular-event reasoning that both human cognition and AI routinely perform. A physician diagnosing a patient is not performing an experiment that will be repeated infinitely; a language model generating a response is not sampling from a repeatable physical trial.

The subjectivist tradition, inaugurated by Ramsey [1] and developed independently by De Finetti [2], treats probability as a coherent degree of belief held by a reasoning agent. On this account, to assign a probability is to make a commitment about one's willingness to act under uncertainty, a betting commitment that must satisfy the axioms of probability theory to avoid guaranteed loss. The philosophical power of this framework is that it makes every rational agent, regardless of their access to frequencies or physical processes, a legitimate holder of probabilistic beliefs. Savage [3] extended the subjective framework into a full theory of rational decision-making under uncertainty, deriving expected utility theory from primitive axioms of preference coherence. The Savage framework [3] is the standard against which both human and machine decision-making is typically evaluated in the economics of choice and the design of rational AI systems.

Cox [17] provided an independent derivation of the probability axioms from purely logical requirements of consistency, establishing that any system of plausible inference that satisfies certain desiderata of common sense must obey the probability calculus. This result, Cox's theorem [17], is philosophically significant because it grounds probability not in any empirical feature of the world but in the logic of rational inference itself. It

implies that a reasoning system need not explicitly represent numerical probabilities in order to be probabilistically rational; what is required is that its beliefs satisfy the constraints of coherent uncertainty management.

The Bayesian epistemology that emerges from this tradition, elaborated philosophically by Bernardo and Smith [18], treats knowledge as the continuous revision of prior beliefs in light of evidence. The prior encodes what the agent knew before evidence arrived; the likelihood encodes how informative the evidence was; and the posterior encodes what the agent should believe given the full conjunction of prior knowledge and new evidence. The philosophical significance of the prior, as Oaksford and Chater [5] emphasize, is that it is not merely a technical starting condition but a representation of the entire accumulated epistemic biography of the agent, everything that shaped its state of knowledge before the current inquiry began.

2.2 | Human Probabilistic Cognition: Approximation, Bias, and Embodiment

The psychological literature on human probabilistic reasoning developed along two major trajectories that have been in productive tension for half a century. The heuristics-and-biases programme of Tversky and Kahneman [19] documented a catalogue of systematic departures from Bayesian rationality in human judgment: The availability heuristic, which substitutes ease of recall for frequency estimation; anchoring, which distorts probability assessments toward irrelevant reference points; and base rate neglect, which allows vivid case-specific evidence to overwhelm general statistical regularities. This research programme, synthesized in the influential account of Kahneman [20], established the dual-process framework in which System 1 fast, automatic, associative cognition is contrasted with System 2 slow, deliberative, rule-governed reasoning, with human probabilistic errors attributed primarily to the dominance of System 1 in conditions where System 2 engagement is required.

The complementary tradition associated with Gigerenzer et al. [21] challenged the normative assumptions of the heuristics-and-biases programme, arguing that human heuristics are not irrational departures from ideal Bayesian inference but ecologically rational strategies that perform well in the environments for which they evolved. The take-the-best heuristic, the recognition heuristic, and the fast-and-frugal decision trees developed in Gigerenzer et al. [21] achieve accuracy comparable to complex Bayesian models in a wide range of natural decision environments while using far less information and computation. This ecological rationality perspective reframes the comparison between human and ideal probabilistic reasoning: the relevant question is not whether human inference matches the Bayesian norm but whether it achieves good outcomes in the environments in which it actually operates.

The neuroscientific dimension of human probabilistic cognition was established by Knill and Pouget [22], who reviewed evidence that the brain represents and manipulates probability distributions at the level of neural population codes. Psychophysical studies of sensory integration indicated that the brain combines information from multiple sensory modalities in ways that closely approximate Bayesian optimal integration, weighting each source by its reliability exactly as a Bayesian ideal reasoner would. Rao and Ballard [23] proposed the predictive coding framework as a neural implementation of Bayesian inference, in which the brain maintains a generative model of its sensory inputs and uses prediction errors, the difference between predicted and actual sensory signals, to update the model continuously. This framework was developed into the free energy principle by Friston [7], who argued that the brain's fundamental computational imperative is to minimize the divergence between its generative model and the sensory evidence it receives.

The role of embodiment in probabilistic cognition was explored by Merleau-Ponty [24], whose phenomenology of perception established that the body is not a passive instrument of the mind but the primary medium through which the world is encountered and interpreted. Contemporary cognitive science, particularly the enactivist tradition of Clark [14], has elaborated the implications of this insight for probabilistic cognition: the priors from which human Bayesian inference begins are not abstract probability distributions but embodied dispositions, sensorimotor expectations, and affective attunements that are as much in the body as in the brain. Damasio [25] provided clinical evidence for the functional role of emotion in

probabilistic decision-making through the somatic marker hypothesis, demonstrating that patients with damage to the ventromedial prefrontal cortex, who retain intact explicit reasoning abilities but lose normal emotional responses, make systematically worse decisions in gambling tasks that require probability learning from feedback.

2.3 | Probabilistic Foundations of Artificial Intelligence

The relationship between probability theory and AI has been deep and generative since the earliest days of the field. Pearl [26] introduced Bayesian networks as a formal representation of probabilistic dependencies among variables, providing AI systems with a principled mechanism for propagating uncertainty through causal structures. The tractability of inference in Bayesian networks, and the limitations encountered in highly connected networks, motivated the development of approximate inference methods, including Monte Carlo sampling and variational techniques, which form the computational backbone of modern probabilistic AI.

The variational inference framework reviewed comprehensively by Blei et al. [10] replaces the intractable computation of exact posterior distributions with an optimization problem: finding the member of a tractable family of distributions that is closest to the true posterior in a specific information-theoretic sense. This approach makes Bayesian inference computationally feasible in high-dimensional models with millions of parameters, and it provides the theoretical foundation for the variational autoencoder and related generative models that achieve state-of-the-art performance in tasks ranging from image synthesis to drug discovery. The information-theoretic quantity that variational inference minimizes, the Kullback-Leibler divergence between the approximate and true posteriors, appears with striking regularity in the comparative analysis of human and artificial cognition: It is the same quantity that measures the departure of human approximate inference from Bayesian optimality, and it is the same quantity that appears in the free energy principle as the cost of maintaining an inaccurate world model.

Ghahramani [8] argued that probabilistic machine learning is not merely a technical methodology but a philosophical stance on the nature of AI: Intelligence is the capacity to represent uncertainty, reason about it coherently, and act wisely under it. This framing directly parallels the Bayesian epistemological tradition in philosophy and creates the conceptual bridge across which the present comparison is conducted. The Gaussian process framework developed in Bishop and Nasrabadi [9] provides a particularly transparent illustration of machine probabilistic reasoning: rather than committing to a single predictive function, the Gaussian process maintains a full probability distribution over functions, expressing genuine uncertainty about which input-output mapping the data supports.

The calibration of machine probabilistic predictions, specifically the alignment between stated confidence and actual accuracy, was examined systematically by Guo et al. [12], who documented a systematic overconfidence in modern deep neural networks and evaluated post-hoc calibration methods, including temperature scaling and Platt scaling. The philosophical significance of calibration extends beyond technical accuracy: A well-calibrated system knows what it does not know, which is a precondition for trustworthy probabilistic advice in any consequential domain.

2.4 | The Comparative Philosophy of Mind and Machine

The philosophical tradition of comparing human and artificial minds has a history coextensive with the development of computing itself. Turing [27] proposed the imitation game as an operational test for machine intelligence, a proposal that implicitly treated intelligence as a probabilistic capacity: the machine succeeds if a human judge, acting on the basis of conversational evidence, cannot reliably distinguish it from a human interlocutor. The philosophical debates provoked by this proposal, including Searle's Chinese Room argument [28] that syntactic symbol manipulation cannot constitute genuine semantic understanding, remain unresolved but have been substantially reframed by the emergence of probabilistic learning systems.

Dreyfus [13] anticipated many of the limitations of rule-based AI that subsequently emerged in practice, arguing that human intelligence is fundamentally grounded in embodied, pre-reflective engagement with the

world rather than in the manipulation of explicit symbolic representations. While the connectionist and probabilistic learning systems that dominate contemporary AI differ fundamentally from the symbolic systems Dreyfus criticized [13], his core philosophical insight, that the informal, contextual, and embodied dimension of human cognition resists formalization without remainder, retains its force and gains new relevance in the present analysis.

The concept of bounded rationality introduced by Simon [29] provided the first systematic philosophical framework for understanding the gap between ideal probabilistic rationality and the constrained reasoning of actual cognitive agents. Simon [29] argued that human beings satisfice rather than optimize, searching for solutions that are good enough rather than globally best, and that this satisficing behavior is not a failure of rationality but a rational response to the computational and informational constraints under which real cognition operates. The connection between bounded rationality and modern approximate inference in machine learning, noted by Gershman et al. [30], suggests that both human and artificial systems are fundamentally in the business of approximating ideal probabilistic inference under resource constraints, but they approximate it in characteristically different ways using characteristically different resources.

The question of machine consciousness, which bears on the comparative analysis in philosophically important ways, was addressed by Nagel [31] in a framework that remains philosophically compelling: The subjective character of experience, what it is like to be a particular kind of mind, cannot be reduced to any objective physical description of the system that instantiates it. Whether AI systems have or could have phenomenal experience remains genuinely open, but Dennett [32] argued that the apparent hard problem of consciousness dissolves under careful heterophenomenological analysis and that the distinction between genuine and mere apparent understanding is less philosophically tractable than it initially seems. The implications of this debate for the comparative analysis of human and machine probabilistic cognition are taken up in Section 5.

3|The Human Mind as Embodied Bayesian Inference

3.1|The Body as a Prerequisite

Human beings do not begin reasoning from a position of epistemic neutrality. The nervous system arrives at birth already equipped with a rich set of prior expectations about the structure of the world, encoded not in propositions but in the firing thresholds of neurons, the gain settings of sensory systems, and the timing properties of neural circuits that reflect the statistical regularities of the species' evolutionary history. The infant who recoils from a looming visual stimulus is exhibiting a prior over the probability that rapidly expanding visual objects signal collision, a prior calibrated over millions of generations of encounters with physical threats. The newborn who preferentially attends to face-like patterns is exhibiting a prior over the probability that face-shaped stimuli are significant, a prior that reflects the social structure of the ancestral environment. As Rao and Ballard [23] showed, even the basic architecture of the visual cortex implements a hierarchical predictive model whose structure encodes statistical regularities of the natural visual world.

What is philosophically significant about this is that the human prior is not a mathematical object chosen for analytical convenience. It is a biological and biographical accumulation: A sediment of evolutionary history, developmental contingency, and personal experience that has physical weight, emotional valence, and resistance to revision proportional to its depth. When Friston [7] argues that the brain's fundamental imperative is to minimize the divergence between its generative model and sensory evidence, he is describing a process that operates on priors of this kind, priors that carry the full biographical specificity of the individual organism. This is what Merleau-Ponty [24] meant by saying that the body is not a possession of the mind but its ground: the prior is not a belief the person holds but a way of being-in-the-world that precedes and conditions all explicit believing.

3.2 | Emotion as Probabilistic Modulation

The dominant philosophical tradition, from Plato through Kant to the early twentieth century, treated emotion as an interference with rational judgment. The probabilistic framework and the neuroscientific evidence assembled by Damasio [25] demand a fundamental revision of this view. Emotions are not disruptions of probabilistic reasoning; they are fast, compressed, and partially implicit probability estimates about the significance of events for the organism's goals and well-being.

Fear is a rapid, pre-reflective update of the probability that the current environment contains a threat, executed by neural circuits that are faster than conscious deliberation and calibrated to the statistical patterns of ancestral danger. The somatic marker hypothesis of Damasio [25] proposes that the body generates anticipatory signals, somatic markers that tag representations of future outcomes with their emotional valence, effectively providing the deliberative system with a rapid, affect-based probability estimate of the desirability of different action options. Patients who lose normal somatic markers through ventromedial prefrontal damage retain the capacity for explicit probabilistic reasoning but make systematically worse decisions in tasks that require learning probability distributions from feedback, providing clinical evidence that emotion is a functional component of human probabilistic reasoning rather than merely an accompaniment to it.

The probabilistic significance of emotion is most clearly visible in the way emotional experience shapes the prior. Kahneman [20] documents the affect heuristic, a pattern in which the emotional valence of an object influences estimates of its risk and benefit in ways that are systematically inconsistent with any independent assessment of the underlying probabilities. On the account developed here, the affect heuristic is not a bias in the pejorative sense but a heuristic use of emotional probability estimates whose calibration may be excellent in evolutionarily familiar domains and poor in novel or statistically complex ones.

3.3 | Rationality Bounded and Ecologically Situated

Simon [29] established the foundational insight that human reasoning is bounded by cognitive capacity, time, and information availability and that this boundedness is not merely a limitation but a structural feature of cognition that generates characteristic patterns of approximate inference. The heuristics documented by Gigerenzer et al. [21] can be understood, within the probabilistic framework, as intelligent responses to the problem of computing posterior distributions under severe resource constraints. The recognition heuristic, which infers that the more familiar option is probably the better one, is a fast and frugal approximation to a Bayesian inference that would require much more information and computation to execute explicitly. Its accuracy in natural environments is a consequence of the fact that familiarity tracks genuine quality in ecological contexts shaped by human experience, making the heuristic an efficient prior.

The ecological situatedness of human probabilistic cognition means that departures from ideal Bayesian norms, which look like irrational biases in the abstract, often reflect rational adaptations to specific environmental structures. A person who overweights vivid personal testimony relative to dry statistical summaries is not making a straightforward cognitive error. They are applying a prior that was calibrated to an ancestral environment in which personal testimony was a more reliable guide to action-relevant probabilities than the kind of numerical abstractions that Tversky and Kahneman [19] used in their laboratory studies. The philosophical point, developed systematically in Oaksford and Chater [5], is that the normative standard against which human probabilistic reasoning is evaluated must be sensitive to the ecological context in which that reasoning operates.

4 | Artificial Intelligence as Explicit Probabilistic Inference

4.1 | The Honest Uncertainty of the Machine

When a well-designed probabilistic AI system outputs a confidence score alongside its prediction, it is doing something that human experts rarely achieve naturally: reporting the full distributional content of its

probabilistic judgment rather than a compressed point estimate. A Gaussian process regression model, of the kind described in Bishop and Nasrabadi [9], does not merely predict the value of an output; it provides a complete probability distribution over possible output values, encoding both the most probable value and the degree of uncertainty about it in a single, mathematically principled representation.

This formal honesty about uncertainty is a genuine philosophical virtue of probabilistic AI. It reflects the capacity of machine systems to maintain exact probability distributions in cases where human cognitive architecture produces only a felt sense of confidence or doubt. The calibration work of Guo et al. [12] showed that this formal honesty does not always translate into empirical calibration; deep neural networks can be wildly overconfident in their probability outputs. But the framework is there, and the tools for achieving good calibration, temperature scaling, Platt scaling, and isotonic regression are available. The machine's honesty about uncertainty is structurally guaranteed by its architecture in a way that human confidence reporting is not.

4.2|The Poverty of the Machine's Prior

The prior in a machine learning system is chosen by its designers for reasons of mathematical tractability and computational convenience rather than epistemic depth. A Gaussian prior over neural network weights reflects the assumption that weights are probably small; it does not reflect any genuine knowledge about the domain the network will learn. A Dirichlet prior over topic distributions in a topic model reflects a mathematical preference for sparse distributions; it encodes nothing about the actual structure of the corpus or the conceptual landscape from which it was drawn.

The poverty of the machine prior is not a limitation that can be overcome by more data, more computation, or more sophisticated algorithms. It is a philosophical consequence of the system's ontology. As Russell [15] argues, the alignment of machine systems with human values requires that those systems have accurate probabilistic models of human preferences, models that can only be constructed from outside the machine's own reasoning apparatus. The machine's prior is always a mathematical imposition from outside, never a biographical accumulation from within. This creates an asymmetry that Dreyfus [13] would have recognized immediately: the machine knows about human experience the way an anthropologist knows about a culture they have only read about, never lived in.

4.3|The Brittleness of Machine Probability Under Distributional Shift

A machine learning system trained on a particular distribution of inputs learns a posterior distribution over outputs that is calibrated to that distribution. When the test distribution diverges from the training distribution, the machine's posterior becomes unreliable in ways that can be catastrophic. Goodfellow et al. [11] documents adversarial examples, inputs that are imperceptibly different from training inputs to human observers but that cause machine classifiers to assign high confidence to wildly incorrect outputs. This phenomenon, which has no clean analogue in normal human cognition, reveals something philosophically important about the structure of machine probabilistic reasoning: the machine's probability assignments are locally consistent within its training distribution but globally fragile in ways that the machine has no mechanism to detect.

The philosophical contrast with human robustness is instructive. Human beings are also susceptible to distributional shift, as Kahneman [20] documents extensively in the domain of novel statistical environments. But human cognitive systems have multiple layers of monitoring that detect when the environment has changed in ways that require a different reasoning strategy: the phenomenology of surprise, the felt sense of unfamiliarity, the emotional signal of disorientation. These signals trigger a qualitatively different mode of reasoning, a heightened engagement of deliberative System 2 processing, that the machine has no access to because it has no phenomenology.

5 | The Uncertainty Signature: A Philosophical Diagnostic

5.1 | Concept and Motivation

The comparative analysis developed in the preceding sections reveals a pattern of structural difference between human and artificial probabilistic reasoning that cannot be captured by any single metric of performance or accuracy. What is needed is a philosophical diagnostic that characterizes the relationship between a reasoning system and uncertainty across multiple dimensions simultaneously. This paper introduces the concept of the Uncertainty Signature for this purpose.

The uncertainty signature of a reasoning system is a structured characterization of how that system relates to the unknown across six philosophically distinct dimensions. These dimensions were identified inductively from the comparative analysis, but they map onto a principled partition of the concept of uncertainty management: the source dimension concerns where uncertainty originates in the system's architecture; the responsiveness dimension concerns how the system updates under evidence; the texture dimension concerns whether uncertainty is represented or experienced; the generativity dimension concerns the system's capacity to expand its own hypothesis space; the calibration dimension concerns the alignment between stated and actual uncertainty; and the self-uncertainty dimension concerns whether the system can be uncertain about the adequacy of its own reasoning apparatus.

Table 1. Comparative uncertainty signatures of human and artificial probabilistic reasoners.

Dimension	Human Reasoner	Artificial Reasoner
Prior source	Embodied biography; evolutionary inheritance; developmental history; cultural inscription; affective weighting; partially inaccessible to introspection	Designer specification; mathematical convenience; training data statistics; fully explicit and revisable; epistemically thin
Belief responsiveness	Bounded and emotionally modulated; contextsensitive; subject to motivated reasoning; strong priors resist revision disproportionate to evidence	Mechanically consistent within training distribution; affect-free; calibration tractable by post-hoc methods; brittle under distributional shift
Texture of uncertainty	Phenomenologically rich; felt and existentially weighted; linked to bodily state and biographical identity; drives qualitatively different reasoning modes	Represented as numerical distribution; structurally external to any self-model; no phenomenological character; cannot trigger qualitative modeswitching
Generativity	High; can construct genuinely novel hypotheses through analogy, metaphor, and counterfactual reasoning; hypothesis space expands throughout life	Bounded by architecture and training distribution; interpolates within the training manifold with high facility; extrapolation beyond distribution is unreliable
Calibration profile	Domain-dependent; well-calibrated in ecologically familiar domains; overconfident in novel statistical environments; susceptible to affect-heuristic distortion	Potentially high precision with adequate training; systematic overconfidence in deep networks; improved by post-hoc methods; calibration does not generalise across distributional shifts
Self-uncertainty	Present; the human reasoner can be uncertain about the reliability and validity of its own priors, about its values, and about the adequacy of its reasoning strategies; foundational to wisdom	Absent in the philosophically meaningful sense; the system has no reflexive access to its own prior structure; cannot question whether its epistemic constitution is adequate to the problem at hand

5.2 | The Self-Uncertainty Dimension and Its Philosophical Significance

The sixth dimension of the Uncertainty Signature, self-uncertainty, deserves extended treatment because it marks what is arguably the deepest philosophical difference between human and machine cognition, one that

has implications for the very concept of wisdom. A human being can be uncertain not merely about the state of the world but about the reliability of the cognitive apparatus through which the world is known. A scientist can question whether their training has produced systematic blind spots. A judge can wonder whether their intuitions about guilt reflect accurate probabilistic judgment or culturally shaped prejudice. A person in grief can recognise that their current assessment of the future is probably distorted by loss, and discount it accordingly.

This reflexive self-uncertainty is what Nagel [31] was pointing toward when he argued that the subjective character of experience is irreducible to objective third-person description: The human reasoner's uncertainty about their own epistemic constitution is an expression of the first-person character of cognition that no external model can fully capture. Dennett [32] would likely respond that this reflexivity is itself a computational process that could, in principle, be implemented in an artificial system. But the important point for present purposes is not metaphysical but functional: Current AI systems do not have this dimension of self-uncertainty, and their absence of it is a philosophically important feature of how they reason, not a mere technical limitation awaiting resolution.

6 | Bias, Failure, and the Parallel Disruptions of Probabilistic Reason

6.1 | Cognitive Biases as Structured Miscalibration

The extensive literature on cognitive biases initiated by Tversky and Kahneman [19] and synthesised by Kahneman [20] can be read, within the probabilistic philosophical framework, not as a catalogue of reasoning defects but as a map of the conditions under which the human Uncertainty Signature loses its calibration. Each major bias corresponds to a specific structural feature of the human probabilistic architecture that generates systematic deviation from Bayesian optimality under particular evidence conditions.

Confirmation bias, the tendency to weight evidence that confirms existing beliefs more heavily than equally informative disconfirming evidence, reflects the dominance of prior investment in belief revision. When a prior is held with high confidence and emotional depth, the Bayesian updating process is distorted in the direction of prior preservation: The effective weight given to disconfirming evidence is reduced, and the posterior moves less than ideal Bayesian inference would demand. The ecological rationality perspective of Gigerenzer et al. [21] notes that this conservatism has adaptive value in stable environments where established beliefs are likely to be approximately correct, and where the cost of belief revision exceeds the benefit of accuracy on any single new piece of evidence.

Base rate neglect, documented experimentally by Tversky and Kahneman [19] and analysed theoretically by Oaksford and Chater [5], reflects the dominance of the experiential prior, the one built from vivid case-specific evidence, over the statistical prior built from abstract numerical information. A person who learns that a friend contracted an illness from a specific source will overestimate their own risk from that source even when the base rate of the illness is extremely low, because the causal narrative provided by the specific story generates a stronger prior update than the abstract statistical fact. This is not irrational within the experiential prior framework; it reflects a weighting scheme calibrated to a world in which causal narratives were more reliably action-guiding than numerical abstractions.

6.2 | Machine Learning Pathologies as Structural Analogues

The machine learning failures documented by Goodfellow et al. [11] and the calibration failures catalogued by Guo et al. [12] are structurally analogous to cognitive biases in ways that the Uncertainty Signature framework makes precise. Overfitting, the failure of a machine model to generalise beyond its training data, corresponds structurally to overconfidence in humans: Both involve excessive precision in the posterior distribution, a failure to maintain the degree of spread that genuine uncertainty about the correct model would require.

A structural parallel: Human confirmation bias and machine overfitting are not merely superficially similar phenomena sharing a descriptive label. They are structurally analogous failures of probabilistic systems operating under different but analogous constraints. In both cases, the system has become excessively confident in a model that was a good approximation under one set of conditions, and has failed to preserve the posterior uncertainty that honest acknowledgment of model inadequacy would demand. The philosophical lesson is the same in both cases: calibration is a continuous discipline, not a one-time achievement, and the most epistemically dangerous system is the one that has stopped questioning the adequacy of its own prior.

Distributional shift failure in machine systems corresponds to the domain-specificity of human heuristic calibration. Both reflect the fact that a probabilistic system whose prior was shaped by one distribution of evidence may perform poorly when the evidence distribution shifts. The difference, philosophically important, is that the human system has phenomenological resources for detecting distributional shift, the felt sense of unfamiliarity, surprise, and disorientation, that trigger mode-switching toward more deliberate reasoning. The machine has no such resources; it applies its training-distribution posterior to out-of-distribution inputs without any internal signal that something has changed.

6.3 | Calibration as Epistemic Virtue

The concept of calibration, the alignment between expressed confidence and empirical frequency of correctness, represents a philosophical ideal that neither human nor machine reasoners achieve completely but that both approach in domain-specific ways. Guo et al. [12] showed that well-trained machine systems can achieve excellent calibration on held-out data from the same distribution as the training set, a genuinely impressive epistemic achievement that human experts rarely match even in their areas of domain competence. However, machine calibration degrades predictably and severely under distributional shift, while human calibration, though more variable across domains, benefits from the phenomenological alarm system that distributional shift triggers.

The philosophical virtue of calibration is not merely technical accuracy. It represents a commitment to what Jaynes and Bretthorst [4] called epistemic humility: the willingness to express the full depth of one's uncertainty rather than compressing it to a confident point estimate. A system, human or machine, that achieves good calibration across the range of its judgments is demonstrating a kind of intellectual integrity that has genuine moral significance when those judgments influence consequential decisions.

7 | Prediction, Surprise, and the Common Grammar of Mind

7.1 | Prediction as the Common Computational Imperative

The most striking finding to emerge from the convergence of neuroscience, cognitive science, and machine learning over the past two decades is that both biological and artificial cognitive systems appear to implement variants of the same fundamental computational strategy: Build a generative model of the environment, generate predictions from it, and update the model in proportion to the error between predictions and observations. This is not a metaphor or a loose analogy. As Friston [7] demonstrated through the free energy principle, the predictive coding architecture of the brain and the variational inference architecture of probabilistic machine learning are mathematically equivalent at the level of the objective they optimise. Both systems minimise a quantity that measures the divergence between their internal model of the world and the evidence the world actually provides.

The predictive coding framework of Rao and Ballard [23] showed that this prediction-error minimisation architecture is implemented at the level of cortical microcircuits, with specific cell populations encoding predictions, specific populations encoding prediction errors, and synaptic connections carrying the learning signal that updates the generative model. The machine learning equivalent, the backpropagation algorithm that updates neural network weights in proportion to prediction error, is formally analogous at the level of

the computational principle it implements, though its biological implementation differs radically from the synaptic mechanisms of cortical learning.

7.2 | The Phenomenology of Surprise and Its Machine Absence

For the human reasoner, the experience of surprise, the moment when the world behaves contrary to a strongly held expectation, is a phenomenologically distinctive event with lasting consequences. It involves a rapid reorientation of attention, a heightened state of arousal, and the initiation of a memory consolidation process that preferentially encodes surprising events for long-term retention. The magnitude of the surprise, in the information-theoretic sense, is directly proportional to the strength of the violated prior. The death of a person close to us, a diagnosis that contradicts an assumption of health, a betrayal from someone trusted: these are experiences of high surprisal whose impact is proportional to the depth and confidence of the priors they violate.

What is philosophically significant is that the phenomenological experience of surprise in human beings is not merely a marker of prediction error but a functional component of the cognitive response to it. The felt quality of surprise triggers a qualitatively different mode of epistemic engagement, a deeper, more reflective reconsideration of the prior that generated the failed prediction, that is not triggered by small, expected prediction errors. Clark [14] argues that this mode-switching capacity, the ability to move from automatic predictive processing to explicit, deliberative model revision in response to sufficiently large prediction errors, is a crucial feature of the human cognitive architecture that has no clear parallel in current machine systems.

For the artificial system, the analogue of surprise is the loss function value: A numerical measure of how different the actual output was from the predicted output. The system updates its parameters in proportion to this numerical quantity but has no phenomenological experience of it, no internal state change that corresponds to the felt quality of surprise, and no mechanism for triggering qualitatively different processing modes in response to large versus small prediction errors. The prediction error is a signal; it is not an event.

8 | Epistemic Ethics: Machine Probability in Human Institutions

8.1 | The Political Ontology of the Machine Prior

When a probabilistic AI system is deployed in a consequential institutional context, a judicial risk assessment tool, a medical diagnostic aid, a financial credit scoring model, a prior is being imposed on the situation that was not chosen by those affected by it. That prior was derived from historical data encoding the patterns of past human decisions, decisions made in social contexts characterised by systematic power differentials, institutional biases, and the accumulated prejudices of the societies that generated them. A machine that learns its prior from this historical record does not transcend that history; it inherits it, and it applies it with the mechanical consistency and formal authority of a probabilistic system whose outputs are typically presented as objective.

Russell [15] argues that the alignment of machine systems with genuine human values requires probabilistic models of human preferences that are far more sophisticated and philosophically transparent than current systems provide. The prior problem in AI ethics is not merely a technical question about which historical data to include or exclude from training; it is a political and philosophical question about whose probabilistic model of human worth, risk, and desert should be encoded in the systems that make consequential judgments about people's lives.

8.2 | Epistemic Colonisation and the Atrophy of Human Judgment

There is a distinctive risk that arises when machine probabilistic confidence is deployed in institutional contexts where human judgment is constitutionally required. The risk, which may be called epistemic colonisation, is the progressive displacement of human deliberative judgment by the authority of machine confidence, not through explicit argument but through the institutional prestige of quantitative precision.

When a judge is presented with a system's seventytwo percent probability that a defendant will reoffend, or when a physician receives a model's eighty-eight percent confidence that a scan shows malignancy, the human decision-maker is placed in an institutionally and psychologically difficult position.

To override the machine's probability estimate requires the human to articulate, in probabilistic terms, the basis for a different posterior, to specify what aspects of the evidence the machine has failed to weight correctly, and to defend this position against the implicit authority of a system that has been trained on more data than any human expert could review. Most practitioners, even highly competent ones, lack the probabilistic training and institutional confidence to do this effectively. The machine's probability report is not a contribution to deliberation; it tends to become the conclusion, clothed in the social authority of numerical objectivity.

The deeper concern, articulated in the framework of Simon [29] and the ecological rationality of Gigerenzer et al. [21], is that systematic deference to machine probability estimates will, over time, atrophy the human capacity for independent probabilistic judgment in precisely those domains where machine confidence is available. A profession that consistently delegates probabilistic reasoning to algorithmic systems will lose, within a generation, the practical knowledge of how to form and defend probabilistic judgments from the ground up. The priors that ground expert intuition, accumulated through years of embodied practice and outcome feedback, are not preserved by formal training; they are preserved by use.

8.3 | Conditions for Genuine Complementarity

The preceding analysis is not an argument against AI or probabilistic machine systems. It is an argument for a philosophically serious account of what kind of human-AI cognitive partnership is genuinely complementary rather than merely substitutive. The Uncertainty Signature analysis reveals that the human and machine profiles are different in ways that are systematically compensatory. The machine has superior computational capacity for maintaining calibrated probability distributions over large hypothesis spaces; the human has superior generative capacity for constructing new hypothesis spaces. The machine has superior formal consistency in belief updating; the human has superior sensitivity to the qualitative character of evidence that resists numerical quantification. The machine has no affective distortion; the human has affective access to the biological and social significance of outcomes that the machine cannot represent.

A genuine human-AI cognitive partnership is not one in which the machine reports probabilities and the human executes them. It is a partnership in which machine precision and human depth are placed in sustained, mutually critical dialogue: The machine's calibrated distributions checking the human's motivated reasoning, and the human's embodied judgment and reflexive self-criticism checking the machine's brittleness and prior poverty.

Designing institutions that realise this complementarity requires, as a minimum, that practitioners be trained to engage critically with machine probability outputs, to identify the cases in which distributional shift renders them unreliable, and to exercise their own probabilistic judgment with confidence rather than deference. It requires that AI systems be designed to express their uncertainty in ways that invite rather than foreclose human deliberation. And it requires a philosophical culture in which the machine's formally honest uncertainty is not mistaken for the richer, self-critical, embodied epistemic virtue that it can approximate but never fully instantiate.

9 | Conclusion

This paper has developed a philosophical framework for understanding both human cognition and AI as species of probabilistic inference that share a common normative standard, Bayesian belief revision, but instantiate it through radically different approximation strategies, prior structures, and phenomenological architectures. The foundational arguments of Ramsey [1], De Finetti [2], and Jaynes and Bretthorst [4] establish probability as the unique consistent framework for rational belief under uncertainty; the neuroscientific research programme of Friston [7] and Rao and Ballard [23] establishes that the brain

implements this framework in the form of hierarchical predictive coding; and the probabilistic machine learning tradition reviewed in Ghahramani [8] and Blei et al. [10] establishes that contemporary AI systems implement it through variational inference and related methods. The convergence is mathematically precise and philosophically important: It means that the comparison between human and machine probabilistic reasoning is not an analogy but a structural analysis of two different implementations of the same computational principle.

The Uncertainty Signature introduced in this paper provides the conceptual vocabulary for conducting that structural analysis with philosophical precision. Across the six dimensions of prior source, belief responsiveness, texture of uncertainty, generativity, calibration, and self-uncertainty, the human and machine profiles reveal a pattern of systematic complementarity that is philosophically more interesting than either a narrative of machine superiority or a narrative of human irreplaceability. The human prior is biographically rich and phenomenologically grounded in ways that no current or foreseeable machine prior can match; the machine's belief updating is formally consistent and affect-free in ways that human cognition, for good reasons rooted in its evolutionary history, is not. The human reasoner can be uncertain about itself; the machine, in any philosophically meaningful sense, cannot.

The cognitive bias literature of Tversky and Kahneman [19] and Kahneman [20] and the machine learning failure modes documented in Goodfellow et al. [11] and Guo et al. [12] are revealed, by the Uncertainty Signature analysis, as structurally analogous disruptions of probabilistic rationality arising from the same root cause: excessive confidence in a model whose prior was calibrated to conditions that may not obtain in the current epistemic environment. The philosophical remedy is the same in both cases, calibration as a continuous discipline, epistemic humility as an active practice, and the maintenance of posterior spread appropriate to the genuine uncertainty of the situation, but the institutional and practical means of achieving it differ between biological and artificial reasoners in ways that matter for the design of human-AI systems.

The ethical and social analysis in Section 8 establishes that the deployment of machine probabilism in human institutions is not an epistemically neutral event. The prior encoded in a machine system is a political object; the confidence it expresses carries institutional authority; and the systematic delegation of probabilistic judgment to algorithmic systems risks atrophying precisely the human cognitive capacities that give probabilistic expertise its depth and practical wisdom its roots. The conditions for genuine human-AI complementarity are demanding: They require practitioners with the training and institutional standing to engage machine probability outputs critically, and they require AI systems designed to invite rather than foreclose human deliberation.

Ultimately, the comparison between human thought and AI conducted in this paper reveals something about the nature of intelligence itself. Any mind, biological or artificial, that engages with an open and uncertain world must take a position on its own uncertainty: it must have a characteristic way of holding what it does not know, revising it honestly, and generating new possibilities from within the constraints of its prior. The human being and the artificial system occupy different, philosophically irreducible positions in the space of possible uncertainty relationships, and the intellectual work of understanding precisely how they differ, and precisely where they can complement each other, is among the most important tasks that the philosophy of mind and the ethics of AI jointly face.

Acknowledgements

The author thanks colleagues at the Department of Statistics, Ravenshaw University, for discussions that sharpened the arguments presented here.

Funding

This research received no external funding.

Data Availability

This is a theoretical article; no datasets were generated or analysed.

Conflict of Interest

The author declares no conflict of interest.

References

[1] Ramsey, F. P., & Braithwaite, R. B. (2000). *The foundations of mathematics and other logical essays*. Routledge. https://books.google.com/books?id=1st-3kYOEPQC

[2] De Finetti, B. (1937). *La prévision: Ses lois logiques, ses sources subjectives*. Institut Henri Poincaré. https://books.google.com/books?id=ofK5OwAACAAJ

[3] Savage, L. J. (1972). *The foundations of statistics*. Dover Publications. https://books.google.com/books?id=zSv6dBWneMEC

[4] Jaynes, E. T., & Bretthorst, G. L. (2003). *Probability theory: The logic of science*. Cambridge University Press. https://books.google.com/books?id=tTN4HuUNXjgC

[5] Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. OUP Oxford. https://books.google.com/books?id=sLetNgiU7ugC

[6] Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788

[7] Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature reviews neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

[8] Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*(7553), 452–459. https://doi.org/10.1038/nature14541

[9] Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning. In *Stat sci* (pp. 140–155). New York: Springer. http://dx.doi.org/10.1117/1.2819119

[10] Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the american statistical association*, *112*(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773

[11] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. https://books.google.com/books?id=-s2MEAAAQBAJ

[12] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 1321–1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html

[13] Dreyfus, H. L. (1972). *What computers can't do: A critique of artificial reason*. Harper & Row. https://books.google.com/books?id=TsraAAAAMAAJ

[14] Clark, A. (2016). *Surfing uncertainty: Prediction, action, and the embodied mind*. Oxford University Press. https://books.google.com/books?id=Yoh2CgAAQBAJ

[15] Russell, S. J. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking. https://books.google.com/books?id=8vm0DwAAQBAJ

[16] Von Mises, R. (2013). *Wahrscheinlichkeit statistik und wahrheit*. Springer Berlin Heidelberg. https://books.google.com/books?id=nuGEBwAAQBAJ

[17] Cox, R. T. (1946). Probability, frequency and reasonable expectation. *American journal of physics*, *14*(1), 1–13. https://doi.org/10.1119/1.1990764

[18] Bernardo, J. M., & Smith, A. F. M. (2009). *Bayesian theory*. Wiley. https://books.google.com/books?id=11nSgIcd7xQC

[19] Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

[20] Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux. https://books.google.com/books?id=ZuKTvERuPG8C

[21] Gigerenzer, G., Todd, P. M., & ABC Research Group, T. (2000). *Simple heuristics that make us smart*. Oxford University Press. https://global.oup.com/academic/product/simple-heuristics-that-make-us-smart-9780195143812?cc=ir&lang=en&

[22] Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in neurosciences*, *27*(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007

[23] Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, *2*(1), 79–87. https://doi.org/10.1038/4580

[24] Merleau-Ponty, M. (1976). *Phénoménologie de la perception*. Gallimard. https://books.google.com/books?id=K_BtPQAACAAJ

[25] Damasio, A. (2008). *Descartes' error: Emotion, reason and the human brain*. Random House. https://books.google.com/books?id=MRY3hmYc1W8C

[26] Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Elsevier Science. https://books.google.com/books?id=AvNID7LyMusC

[27] Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

[28] Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

[29] Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, *69*(1), 99–118. https://doi.org/10.2307/1884852

[30] Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278. https://doi.org/10.1126/science.aac6076

[31] Nagel, T. (1974). What is it like to be a bat? *The philosophical review*, *83*(4), 435–450. https://doi.org/10.2307/2183914

[32] Dennett, D. C. (1991). *Consciousness explained*. Little, Brown and Company. https://books.google.com/books?id=gpncAAAAIAAJ